Probabilistic graphical models for statistical genetics and survival analysis. Application to the Lynch syndrome

Alexandra Lefebvre

SORBONNE UNIVERSITÉ

Laboratoire de Probabilités Statistique et Modélisation - LPSM, CNRS 8001

École doctorale 386 : Sciences Mathématiques de Paris Centre

THÈSE DE DOCTORAT

pour obtenir le grade de

DOCTEUR EN MATHÉMATIQUES

Spécialité : Mathématiques Appliquées

Présentée par

Alexandra LEFEBVRE

Probabilistic graphical models for statistical genetics and survival analysis. Application to the Lynch syndrome.

---

sous la direction de Grégory NUEL et Patrick BENUSIGLIO

Soutenue le 17 janvier 2022 devant la commission d'examen composée de

Adeline Leclerc Samson	Université de Grenoble	Rapporteur
Simon de Givry	INRAE MIAT	Rapporteur
Dominique Stoppa-Lyonnet	Institut Curie	Examinateur
Stéphane Robin	Sorbonne Université	Président du jury
Grégory Nuel	Sorbonne Université	Directeur de thèse
Patrick Benusiglio	APHP Sorbonne Université	Co-directeur de thèse

Remerciements

Je tiens tout d'abord à remercier chaleureusement mes directeurs, Grégory et Patrick, pour la confiance qu'ils m'ont accordée, leur soutien indéfectible et leur enthousiasme contagieux. Merci Grégory pour ta générosité, ton imperturbable foi en les profils bi-disciplinaires et ton insatiable curiosité. Merci Patrick pour ton accompagnement bienveillant dans la démarche clinique, tes efforts pour comprendre mon jargon mathématique et ton enthousiasme pour me convaincre de l'utilité du modèle développé dans ma thèse pour les cliniciens et la nécessité de le voir aboutir. Merci à tous les deux pour m'avoir transmis votre expérience avec tant de patience et de bienveillance.

Je tiens également à exprimer ma gratitude à mes deux rapporteurs, Adeline Leclercq Samson et Simon de Givry pour le travail qu'ils m'ont accordé (d'autant plus au vu du nombre de pages) et leurs précieux conseils.

Un grand merci également à mes deux examinateurs, Dominique Stoppa-Lyonnet et Stéphane Robin, pour l'honneur qu'ils m'ont fait de participer à mon jury, le temps qu'ils m'ont accordé et les discussions qui s'ensuivirent pour élargir ce travail.

Je tiens à exprimer ma profonde reconnaissance à la Ligue contre le cancer qui a financé ma thèse. J'admire la passion des gens qui la compose et je remercie chaque personne qui contribue à ses engagements. Mes pensées vont vers Axel Kahn et ses proches.

Un grand merci à Alex Duval pour m'avoir guidée lors des premiers mois, fait découvrir avec tant de patience la richesse de la biologie de mon sujet et permis de rencontrer les praticiens hospitaliers qui m'ont accompagnée dans mon travail.

Je souhaiterais remercier l'ensemble des personnes du LPSM, mon laboratoire d'accueil. Un immense merci tout d'abord à mes camarades de premier bureau : Eric, Alice, Omar, Saad, Yi puis Lucas I., Othman, Jean-David et ceux de mon dernier bureau, Vivien, William, Alexandre, Malo, Michel, Chenguang, Sandro puis Lucas D., Jérôme, Bastien, Sergi, Edhin. Et bien sûr les doctorants / postdoctorants avec qui je n'ai pas eu la chance de partager un bureau, ni d'Avalon jusqu'ici, mais tant de rires et de soutien, vous avez pimenté mon expérience pendant ces quatre années. Merci aux plus anciens: Paul M., Carlos B., Henri E., aux moins anciens: Vivien, Flaminia, Adeline, Florian, Isao, Nicolas M., Nicolas G., Sebastien F., Laure M., Barbara D., Simon C., Guillaume C., Clément C., Fabio C. et aux nouvelles

recrues : Antonio, Yoan, Lucas B., Emilien, David, Loïc, Robin, Guillaume, Ikraa, Ariane, Pierre, Ludovic, Francesco, Miguel. Merci particulièrement à Eric et Vivien pour m'avoir accompagnée dans mes premiers pas en thèse et jusque dans les derniers et après. Merci Flaminia pour ta joie de vivre, merci Adeline pour ton soutien à toute épreuve. J'espère que nos amitiés nous emmèneront loin.

Un grand merci au secrétariat, Florence, Serena, Josette, Elise, Fatime, Louise, Valérie, Nathalie, Corinne pour votre aide précieuse dans toutes les démarches administratives. Et j'en profite pour te remercier, Hugues, pour m'avoir permis d'éviter de nombreux désagréments informatiques.

Merci Stéphane, Jean-Noël, Gauthier, Charlotte, Emeline, Maud, Antonio, Chloé, Nina avec qui j'ai enseigné et merci à mes étudiants qui m'ont fait découvrir la joie de transmettre. Merci Amaury, Anna, Stéphane pour vos conseils et pour m'avoir souvent proposé votre aide.

Merci aux personnes avec qui j'ai découvert le plaisir de collaborer pour un projet commun, Olivier Bouaziz, Sabine Mercier, Vittorio Perduca. Merci d'avoir attrapé ces opportunité d'aventure ensemble, j'espère qu'elles se poursuivront et j'en profite pour remercier Florence Coulet et Erell Guillerm qui vont monter dans le train sur un prochain sujet. Merci également à l'équipe SUMMIT (Carnot) et en particulier celles et ceux avec qui j'ai travaillé, Thomas, Kaichen ainsi que Marie, Nora et un grand merci à France Active pour leur confiance et pour tous nos échanges.

Un grand merci à l'équipe de génétique de l'institut Curie et en particulier Dominique Stoppa-Lyonnet, Antoine de Pauw, Marine le Mentec et Anaïs pour votre chaleureux accueil et pour avoir guider mes premiers pas dans l'univers du conseil en génétique.

Je suis tout particulièrement reconnaissante envers les personnes responsable de la formation continue de Sorbonne Université et de Paris-Saclay pour m'avoir soutenue dans ma volonté de reprise d'études. Un merci particulier à mes professeurs qui n'ont pas douté qu'il est possible de se réorienter dans les mathématiques même tardivement. Un nouveau merci à Adeline Leclercq Samson qui a su répondre, il y a quelques années, avec beaucoup de patience et de détails, à l'email parsemé d'interrogations d'une inconnue souhaitant reprendre les mathématiques appliquées à la biologie sans savoir comment remettre un pied dedans.

Merci à toutes les personnes qui ont rendu les retraites scientifiques si enthousiasmantes, Grégory, Flora, Sabine, Olivier, Vittorio, Vivien, Allan. Merci pour nos nombreux échanges et discussions qui font naitre les envies et les idées.

Un grand merci à mes ami.e.s pour être à mes côté, pour tous nos moments partagés et pour votre soutien ... Un merci tout particulier à Céline, Jo, Mélinée, Robin, Hélène, Christophe, Philippe et Dominique pour avoir tant cru en moi et pour m'avoir en plus permis d'améliorer mes conditions de rédaction en plein covid.

Merci mon Marco pour toutes les surprises qui jonchent nos rencontres! Et merci Nico L., Marjo, Nico M., Audrey, Gillou, Sandie, Marie, Aurélie N., Oriane, Martina, Florence D., Jeanne, Stéphane, Jake, Pradeep. Merci à ceux que la vie a éloignés géographiquement mais que je suis sûre de retrouver, Michèle, Dino, la p'tite Flo, Valoche, Aurélie B., Alberto, Nath., Philippe R., Jenny, Nadège, André, Serge, Seb.

Je remercie bien sûr mes parents et mes grands-parents, en particulier ma mère qui me donne la force de croire que tout est possible pourvu que ce soit fait avec passion. Et merci à ma famille, mes frère et soeur, mes oncles, tantes, cousins et cousines dont l'amour et le soutien ont tant contribué.

Je ne pourrai être exhaustive, ces lignes omettent tant des personnes. Je remercie toutes celles et ceux qui ont illuminé ma vie.

Résumé

Cette thèse porte sur l'inférence exacte dans les réseaux bayésiens et les chaînes de Markov cachées (HMMs) et ses applications en survie multi-états, génétique et segmentation. Elle est financée par *la Ligue Contre le Cancer*[1] (LNCC) et un de ses principaux objectifs est le développement d'un modèle de calcul de risque de prédisposition génétique et risque de cancer dans le cadre du syndrome de Lynch. Elle comporte deux grands axes que sont 1) L'extension de l'algorithme somme-produit sur l'anneau polynomial et application en segmentation et génétique familiale; 2) L'intégration de modèles de survie multi-états aux modèles de génétique familiale pour le calcul de risques en oncogénétique avec localisations et diagnostics multiples.

Les réseaux bayésiens et les chaînes de Markov cachées sont omniprésents dans les applications biomédicales et plus généralement en biologie en ce sens qu'ils permettent de modéliser les relations probabilistes entre des variables latentes et des variables observées (noeuds des graphes) régies par une structure de dépendance particulière (arêtes des graphes). Cette structure de dépendance est exploitée pour réduire la complexité algorithmique d'une inférence exacte dans ces graphes. L'algorithme *somme-produit*, aussi appelé *message-passing* ou *belief propagation* dans les réseaux bayésiens et *forward-backward* dans les HMMs, permet de réduire cette complexité de exponentielle avec le nombre de variables à exponentielle avec la largeur arborescente du graphe dans un réseau bayésien et linéaire avec le nombre de variables dans une HMM.

Les réseaux bayésiens sont particulièrement appropriés à l'études des maladies multifactorielles avec une composante génétique en présence de données phénotypiques familiales car ils permettent de modéliser la structure de dépendance entres les génotypes (la plupart latents) des membres de la famille et les phénotypes (le plus souvent observés). Inférer la loi des génotypes par la force brute deviendrait rapidement infaisable même avec de petites familles sans les algorithmes existants. Les réseaux bayésiens appliqués à le génétique familiale sont appelés communément *pedigree-based models*.

Un travail de Master 2 préliminaire consistant à implémenter un modèle simple de calcul de risque appelé le modèle de Claus-Easton (Claus et al., 1991; Easton et al., 1993) et composé d'un gène majeur à transmission autosomique dominante et une maladie, le cancer du sein, nous a amenés, au file des collaborations, à constater deux choses: 1) Les *pedigree-based models* actuels ne proposent pas le calcul de risques familiaux notamment la distribution du nombre de porteurs d'un variant (ou allèle)

[1] `https://www.ligue-cancer.net`

délétère dans une famille qu'une extension simple et déjà existante de l'algorithme permettrait d'obtenir; 2) Les modèles mathématiques de calcul de risque dans le cadre du syndrome sein/ovaire sont multiples et régulièrement mis à jour. Pour le syndrome de Lynch de tels modèles sont rares et peu adaptés à l'évolution des données biologiques. Ces deux simples constats ont été les points de départ des deux grands axes de ma thèse:

- Extension de l'algorithme somme-produit sur l'anneau polynomial. Lors de l'implémentation d'une version à potentiels polynomiaux pour le calcul de fonctions génératrices des probabilités du nombre de porteurs dans une famille, nous avons réfléchi à des extensions plus variées sur l'anneau polynomial avec notamment, par exemple, le calcul des dérivées de la vraisemblance dans un réseau bayésien paramétrique jusqu'à un ordre choisi. D'autre part, suite à une collaboration avec des collègues du domaine de la segmentation, nous avons constaté que l'emploi des fonctions génératrices des probabilités dans le monde de la segmentation peut être particulièrement intéressant pour relaxer la contrainte du prior sur l'espace des segmentations (incluant le nombre de segments) des méthodes actuelles.

- Développement d'un modèle de calcul de risques dans le cadre du syndrome de Lynch, une prédisposition génétique au cancer. Le spectre de Lynch étant très large (colon, rectum, endomètre, ovaire, estomac, intestin grêle, voies biliaires, pancréas, uretère, rein, vessie, prostate, etc.) et les récurrences et événements multiples chez un même individu n'étant pas rare, un modèle de survie multi-états s'impose.

Cette thèse est divisée and trois parties et huit chapitres non-indépendants, à l'exception du chapitre 6 qui peut être lu indépendamment.

Partie I. La première partie est essentiellement constituée d'introductions (à une exception près) aux différents domaines socles de la thèse.

Le chapitre 1 est une introduction aux réseaux bayésiens et à l'algorithme somme-produit pour l'inférence exacte sur ces réseaux. Il contient également une section dédiée aux cas particuliers des HMMs et des chaînes de Markov. Il est le fruit d'un état de l'art et d'une compréhension des outils existants afin d'en proposer une synthèse. Ces outils constituent le socle de toutes les méthodes et applications développées par la suite.

Le chapitre 2 propose une introduction à l'analyse de survie et en particulier aux modèles multi-états dans ce domaine. Dans notre contexte, les modèles multi-états permettent de modéliser l'évolution d'un individu dans le temps à travers différents états (sain, diagnostiqué avec la maladie 1, diagnostiqué avec la maladie 2, etc.). Une transition d'un état vers un autre est appelée un événement. L'enjeu principal est l'estimation des probabilités de transition. En fin de chapitre, nous développons une contribution dans ce domaine avec une version HMM et une discrétisation du temps

afin de simplifier les calculs de ces probabilités au prix d'une hypothèse supplémentaire selon laquelle deux événements ne peuvent pas se produire dans un même pas de temps.

Le chapitre 3 commence par la reprise des éléments de base de biologie moléculaire essentiels à la compréhension de la notion de prédisposition génétique puis explique comment la combinaison d'outils vus dans les chapitres 1 et 2 permet de construire les modèles appelés *pedigree-based models* omniprésents pour l'estimation de paramètres en épidémiologie génétique et l'estimation de risques en conseil génétique. Les principaux algorithmes qui relient théorie des graphes et inférence dans les *pedigree-based* models sont rappelés en fin de chapitre.

Partie II. La deuxième partie est consacrée au premier axe de la thèse, à savoir, le développement méthodologique de l'algorithme somme-produit sur l'anneau polynomial et applications.

Le chapitre 4 constitue une brève introduction à l'utilisation des fonctions génératrices dans l'algorithme somme-produit afin de calculer diverses quantités d'intérêt et en particulier la distribution et les moments d'un nombre d'événements dans les réseaux bayésiens. C'est une application directe de méthodes déjà développées, particulièrement dans le contexte des HMMs.

Le chapitre 5 constitue une des premières contributions de cette thèse avec l'extension des méthodes fondées sur les fonctions génératrices au calcul des dérivées de la vraisemblance dans les réseaux bayésiens.

Le chapitre 6 part du constat développé par Mercier and Nuel (2021) de l'existence d'une relation duale entre les méthodes fondées respectivement sur la statistique du score locale et sur les HMMs pour la recherche d'un segment atypique dans une séquence. Aucune méthode non-supervisée ne permettant actuellement d'estimer une fonction de score pour le score local, nous avons alors voulu implémenter une méthode simplement fondée sur un algorithm EM dans une HMM contrainte pour proposer un tel outil. Une initialization correcte de l'algorithme est indispensable pour imposer un prior $\{N = 1\}$ ou $\{N = 0\}$ segment dans la séquence. Plus généralement, nous proposons ensuite une extension des méthodes dites *segment-based* cherchant à détecter des ruptures dans les séquences, à un prior arbitraire sur le nombre de segments et nous montrons qu'une approche fondée sur des potentiels polynomiaux permet de réduire la complexité des calculs.

Partie III. La troisième partie est consacrée au développement du modèle de calcul de risque de prédisposition génétique et risque de cancer dans le cadre du syndrome de Lynch. Le syndrome de Lynch est une prédisposition génétique définie par une mutation pathogène monoallélique constitutionnelle dans un gène du système de réparation des mésappariements de l'ADN. Quatre gènes principaux sont impliqués. Le syndrome de Lynch touche 0.36 % de la population générale, le rendant au moins aussi fréquent que les mutations délétères dans les gènes BRCA1 et BRCA2 impliqués dans le syndrome sein/ovaire. Il confère un risque accru de cancer dans un spectre très large, à savoir, colon, rectum et endomètre principalement mais aussi ovaire, estomac, intestin grêle, voies biliaires, pancréas, uretère, rein, vessie, prostate. Un

homme porteur de ce syndrome a un risque de 55% de développer un cancer colorectal et 70% toutes localisations confondues avant 75 ans. Une femme porteuse de ce syndrome a un risque de 50% de développer un cancer colorectal, 50% un cancer de l'endomètre et 84% toutes localisations confondues avant 75 ans. L'estimation du risque d'être porteur de ce syndrome par un onco-généticien ou un conseiller en génétique est essentiel pour l'orientation vers des tests génétiques et/ou une adaptation appropriée de la surveillance des patients et des membres de leur famille. Cette estimation se fait conditionnellement à l'histoire familiale de cancer et diverses covariables (en particulier des tests biologiques sur tumeur). Le déploiement de ces tests depuis le *universal screening of Lynch syndrome* recommandé pour tous les cancers colorectaux et endometriaux diagnostiqués avant 70 ans (Vasen et al., 2013) aboutit à une plus grande disponibilité de ces données. D'autre part, l'accessibilité accrue des outils de *next generation sequencing* engendre une accumulation de variants (ou allèles) séquencés de pathogénicité inconnue. L'analyse jointe des données et notamment d'une histoire familiale requiert de plus en plus le recours à des outils mathématiques pour accompagner les cliniciens dans leurs prises de décisions. La modélisation des histoires personnelles de cancer dans un modèle multi-états à plusieurs états transitoires pour le syndrome de Lynch nous semble indispensable au vue de l'étendue de son spectre et de la fréquence des événements multiples.

Le chapitre 7 est introductif et retrace les principales notions cliniques indispensables à la compréhension du modèle, notamment l'épidémiologie du syndrome de Lynch, ses manifestations cliniques, les tests existants pour sa détection ainsi que les principaux outils actuels pour l'estimation de risques associés à ce syndrome.

Le chapitre 8 détaille la construction du modèle et offre une sélection d'exemples simulés pour montrer ses atouts et inconvénients par rapport aux modèles existants. Il s'ouvre également sur une discussion quant aux perspectives envisagées.

Abstract

This PhD thesis deals with exact inference in Bayesian networks and hidden Markov models (HMMs) and their applications in multi-state survival, genetics and segmentation. It is funded by *la Ligue Contre le Cancer*[2] (LNCC) and one of its main goal is the development a model which computes risks of genetic predisposition and cancer risks in the framework of the Lynch syndrome. Two main themes are explored: 1) Extensions of the sum-product algorithm on the polynomial ring and application to segmentation and familial genetics; 2) Integration of multi-state survival models in familial genetic models for computing risks in cancer genetics with multiple localizations and multiple diagnosis.

Bayesian networks and Hidden Markov Models are omnipresent in biomedical areas as they allow for modeling probabilistic relationships between latent variables and observed variables (nodes of graphs) linked by a particular structure dependency (edges of graphs). That structure dependency is exploited to reduce the time complexity for an exact inference in these graphs. The *sum-product* algorithm, also called *message-passing* or *belief-propagation* algorithm in Bayesian networks and *forward-backward* algorithm in HMMs leads to a time complexity reduction from exponential in the number of variables to exponential in the treewidth of the graph (respectively linear in the number of variables) in Bayesian networks (respectively in HMMs).

Bayesian networks are particularly suited for studying multifactorial diseases with a genetic component in the present of familial phenotypic data as they allow for modeling the structure dependency between genotypes (usually latent) and phenotypes (usually observed) of family members. The inference of the marginal distribution of a genotype using brut force would rapidly become intractable even in small families without existing algorithms. Bayesian networks applied to familial genetics are usually called *pedigree-based models*.

A previous work during my Master 2 internship consisted in implementing a simple risk prediction model in the framework of the breast cancer named the Claus-Easton model (Claus et al., 1991; Easton et al., 1993) and composed of a single major gene with an autosomic dominant mode of inheritance and a single disease (breast cancer). That work, along fructuous collaborations, lead us to realise two facts: 1) Current *pedigree-based models* do not propose familial risks computations and in particular the distribution of the number of carriers of deleterious variants (or alleles) in a family whereas a simple and existing extension of current algorithms would allow such computations; 2) Mathematical models for breast/ovarian syndrome are

[2] `https://www.ligue-cancer.net`

numerous and regularly updated whereas they are rare and not well suited to the evolution of biological data in the framework of the Lynch syndrome.

These two noticeable facts were starting points for the main two directions of the thesis:

- Extensions of the sum-product algorithm on the polynomial ring. Implementing a version with polynomial potentials for computing the probability generating function of the number of carriers in a family led us to think of more various extensions on the polynomial ring with, for instance, the computation of the derivatives of the likelihood in parametric Bayesian networks up to a chosen order. Moreover a collaboration with colleagues in the field of segmentation led us to notice the fact that generating functions in segmentation could be of particular interest for relaxing the constraint on the prior for the segmentation space (including the number of segments) in current methods.

- Development of a model for computing risks in the framework of the Lynch syndrome, a genetic predisposition to cancer. The Lynch spectrum being wide (colon, rectum, endometrium, ovary, stomach, small bowel, biliary tract, pancreas, uereter, kidney, gallbladder, prostate, etc.) and recurrences and multiple events being not rare, a multi-state survival model is essential.

This thesis is divided into three parts and eight non-independent chapters except Chapter 6 which can be read independently.

Part I. The first part is mostly composed of introductions (except one point) to foundation fields of the thesis.

Chapter 1 is an introduction to Bayesian networks and the sum-product algorithm for exact inference in these models. It also contains a section dedicated to the particular cases of HMMs and Markov chains. It is the result of a state of art and understanding of existing tools in order to propose a summarized version. These tools constitute the foundations of all methods and applications developed during the thesis.

Chapter 2 proposes an introduction to time-to-event data analysis and in particular multi-state survival models. In our framework, multi-state models allow for modeling the evolution of a patient through time via different states (healthy, diagnosed with disease 1, diagnosed with disease 2, etc.). A transition from a state to another one is called an event. The main concern is the estimation of transition probabilities. At the end of the chapter we develop a contribution in this field with an HMM and a time discretization in order to simplify computations at a cost of an additional hypothesis which stands that no more than one event can occur in a time step.

Chapter 3 starts with essential notions in molecular biology for understanding genetic predisposition. Then we pursue with explanations on how the combination of tools seen in Chapter 1 and 2 leads to building models called *pedigree-based models*

which are omnipresent for parameter estimation in genetic epidemiology and risks assessment in genetic counseling. Main algorithms which link inference in graphs and *pedigree-based* models are given at the end of the chapter.

Part II. The second part of the thesis is dedicated to the first direction of the thesis, i.e. methodological extensions of the sum product algorithm on the polynomial ring and applications.

Chapter 4 is a brief introduction to the use of generating functions in the sum-product algorithm in order to compute various quantities of interest and in particular the distribution and moments of a number of events in Bayesian networks. This is a direct application of existing tools in particular in HMMs.

Chapter 5 constitute the first main contribution of this thesis with extensions of methods based on generating functions for the computation of the derivatives of the likelihood in Bayesian networks.

Chapitre 6 is motivated by the work of Mercier and Nuel (2021) who proved a dual relation between methods based on the local score and HMM based methods for the search of an atypical segment in a sequence under certain conditions. To the best of our knowledge there exists no method for the unsupervised learning of a scoring function for the local score. Therefore we decided to develop a simple method based on the EM algorithm in a constrained HMM in order to propose such a tool. A proper initialization of the algorithm is essential for constraining $\{N = 1\}$ or $\{N = 0\}$ segment as prior number of segments in the sequence. We pursue this work with an extension applicable to segment-based methods for allowing for an arbitrary prior number of segments. We show that a version with polynomial potentials allows for time complexity reduction.

Part III. The third part is dedicated to the development of the model for computing risks of genetic predisposition and cancer risks in the framework of the Lynch syndrome. The Lynch syndrome is a genetic predisposition defined as a pathogenic mono-allelic mutation in a gene involved in the mismatch repair system. Four main genes are implicated. Lynch syndrome affects 0.36% of the general population which renders it at least as frequent as pathogenic mutations in BRAC1 BRCA2 involved mostly in breast/ovarian cancer. It confers an increased risk of cancer in a wide spectrum (colon, rectum and endometrium mainly but also ovary, stomach, small bowel, biliary tract, pancreas, ureter, kidney, gallbladder, prostate. A male carrying that syndrome has a 50% risk of developing a colorectal cancer and a 70% risk of developing any cancer in the Lynch spectrum by age 75. A woman carrying that syndrome has a 50% risk of developing a colorectal cancer, a 50% risk of developing an endometrial cancer and a 84% risk of developing any cancer in the Lynch spectrum by age 75. Assessing risks of Lynch syndrome is essential for clinicians in order to adapt germline screening prescriptions and surveillance of patients and their family members. Such estimation is done conditional on a family history of cancer and various covariates (in particular results of biological tests on a tumor). The spreading of these tests since the *universal screening of Lynch syndrome*, recommended for all colorectal and endometrial cancer diagnosed before age 70 (Vasen et al., 2013), leads

to an augmented availability of such data. Moreover the increased access to *next generation sequencing* leads to an accumulation of sequenced variants (or alleles) of unknown pathogenicity. The joint analysis of these data and in particular the family history more and more requires the use of mathematical models to support clinicians in their decision making. Modeling personal histories of cancer in a multistate model with several transient states for the Lynch syndrome is essential for its spectrum to be wide and multiple diagnosis for carriers of the syndrome to be not rare.

Chapter 7 is an introductory chapter and draws main clinical notions essential for understanding the construction of the model. In particular we detail the epidemiology of the syndrome, its clinical and molecular aspects, existing biological testing as well as main existing mathematical tools.

In Chapter 8 we detail the construction of the model and propose a selection of simulated examples to show its main advantages and disadvantages compared to existing models. We conclude with a discussion on perspectives considered.

Contributions

The scientific contributions of my PhD thesis can be divided into two main themes:

- Sum-product algorithm on the polynomial ring and discrete mathematics and applications (1 published article, 6 contributed talks, 2 unsubmitted articles).
- Sum-product algorithm and survival analysis with applications in familial genetics and the Lynch syndrome (1 published article, 1 published opinion paper, 3 contributed talks, 3 posters, 1 unsubmitted article).

Thereafter, the detailed list of published articles and forthcoming articles along with seminars, committees, symposiums, conferences (talks and posters).

Sum-product algorithm on the polynomial ring and discrete mathematics and applications

Computing derivatives in Bayesian networks

- **Published article**: *A sum-product algorithm with polynomials for computing exact derivatives of the likelihood in Bayesian networks,* Lefebvre A. and Nuel G., Proceedings of Machine Learning Research, Vol. 72, p. 201-212, 2018.
- **Invitations** (1 seminar): *Mathematics for Biology seminar*, Institut de mathématiques de Toulouse, France (2019).
- **Conferences** (4 contributed talks): *The International Conference on Probabilistic Graphical Models*, (PGM) 2018, Budapest, Czech Republic. *International Workshop on Applied Probability*, (IWAP) 2018, Prague, Hungary. *Journés Francophones sur les Réseaux Bayésiens et les Modèles Graphiques Probabilistes*, (JFRB) 2018, INRA, Toulouse, France. *European Mathematical Genetics Meeting*, (EMGM) 2018, Cagliari, Italy.

Unsupervised score learning in segmentation

- **Unsubmitted article**: *Unsupervised learning of a scoring function for the maximal scoring segment of a sequence*, Lefebvre A., Mercier S. and Nuel G.
- **Conferences** (1 contributed talk): *Statistical Methods for Post Genomic Data* (SMPGD) 2019, Barcelona, Spain.

Posterior distribution of the number of carriers in familial genetics

- **Unsubmitted article**: *Computing posterior familial risks and distribution of the number of carriers in genetic diseases*, Lefebvre A. and Nuel G.

- **Conferences** (1 contributed talk): *European Mathematical Genetics Meeting* (EMGM) 2019, Dublin, Ireland.

Sum-product algorithm and survival analysis with applications to familial genetics

Computing competing risks in familial genetic models

- **M2 published article**: *Computing individual risks based on family history in genetic diseases in the presence of competing risks,* Nuel G., Lefebvre A. and Bouaziz O., Computational and Mathematical Methods in Medicine, article ID 9193630,Vol. 2017.

- **Conferences** (2 contributed talks, 1 poster): *Statistical Methods for Post Genomic Data*, (SMPGD) 2018, Montpellier, France (talk). *Assises de génétique humaine et médicale*, 2018, Nantes, France (poster). *Statistical Analysis of Multi-Outcome Data*, (SAM) 2017, Liverpool, United Kingdom (talk).

Lynch syndrome

- **Published opinion paper**: *Overcoming the challenges associated with universal screening for Lynch syndrome in colorectal and endometrial cancer,* Benusiglio P., Coulet F., Lefebvre A., Duval A. and Nuel G., Genetics in Medicine, Vol. 22, p. 1422-1423, 2020. I had a light role for this paper.

- **Unsubmitted article**: *LynchRisk: a pedigree-based model for risks computations in the framework of the Lynch syndrome*, Lefebvre A., Benusiglio P., Coulet F., Guillerm E., Duval A. and Nuel G.

- **Invitations** (1 symposium, 1 seminar, 2 conferences & 1 medical committee) *AI and data science for biology*, i-bio & SCAI, Sorbonne Université, Paris, France (symposium, 2021). *Interdisciplinary seminar*, Institut supérieur du calcul et des données (ISCD), Sorbonne Université, online (seminar, 2021). *Young Statisticians and Probabilists* (YSP), online (conference, 2021). *Interplay between Oncology, Mathematics and Numerics* (IbOMaN), online (conference, 2020), *Comité Oncogénétique des tumeurs Digestives de l'Ile-de-France* (CODIF), Institut Curie, Paris, France (medical committee, 2020).

- **Conferences** (1 contributed talk, 1 oral flash poster presentation, 2 posters): *Mathematics of Complex Systems in Biology and Medicine,* Centre International de Rencontres Mathématiques (CIRM) 2020, Marseille, France (talk). *Assises de génétique humaine et médicale* 2020, Tours, France (oral flash poster). *International Society for Gastrointestinal Hereditary Tumors biennial meeting*

(InSiGHT) 2019, Auckland, New Zealand (poster, not presenter). *International Biometric Conference* (IBC) 2018, Barcelona, Spain (poster).

Abbreviations (in alphabetic order)

Abbreviations related to graphs

- BN: Bayesian Network
- DAG: Directed Acyclic Graph
- dMM: discretized Markov Model
- HMM: Hidden Markov Model
- LPD: Local Probability Distribution

Abbreviations related to biology and clinic

- CMMRD : Constitutional MisMatch Repair Deficiency
- GP: General Population
- IHC: ImmunoHistoChemistry
- LS: Lynch Syndrome
- MMR: MisMatch Repair
- MSI: MicroSatellite Instability
  - MSI-H: MSI-high
  - MSI-L: MSI-low
  - MSS: MicroSatellite Stable
- NGS: Next Generation Sequencing
- PCR: Polymerase Chain Reaction
- TM: TransMembrane
- VUS : Variant of Unknown (or Uncertain) Significance

Abbreviations related to diseases

- CC: Colon Cancer
- CRC: ColoRectal Cancer
- EC: Endometrial Cancer
- GIC: upper GastroIntestinal Cancer
- OC: Ovarian Cancer
- RC: Rectal Cancer
- UC: Urinary tract Cancer

Contents

Part I

Introduction

Chapter 1

Introduction to belief propagation in probabilistic graphical models

Sommaire

This introductory chapter is the result of an in-depth review of the literature on probabilistic graphical models and in particular Bayesian networks (BNs). It contains a

summary of a state-of-art related to exact inference in these models and constitute the backbone of all methods and applications later developed in the thesis. This chapter was deeply inspired by Pearl (1986, 1988); Lauritzen and Spiegelhalter (1988); Shafer and Shenoy (1990); Lauritzen (1992, 1996); Cowell et al. (1999); Jensen and Nielsen (2007); Koller and Friedman (2009) among other references.

Probabilistic graphical models play a central role for reasoning in complexe systems involving latent variables. They are omnipresent in a broad range of fields including statistical physics, medical diagnosis, familial genetics, gene networks, speech recognition, etc. Their application to genetics goes back to Wright (1921, 1934) who studied heritable properties of natural species. Probabilistic graphical models provide a graphical representation of the dependency structure in a joint distribution. They mostly involve three types of questions: structure learning, inference and parameter estimation. For the last two, the graph structure is exploited to reduce the computational complexity of an inference or parameter estimation. We propose in this chapter a step by step introduction to main foundations of algorithms involved in this complexity reduction adopting both a theoretical and practical point of view with a particular focus on the *sum-product* algorithm, also called *message-passing* or *belief propagation* algorithm.

This chapter is organized as follows: we start in Section 1.1 with principal definitions and properties in BNs as well as a brief introduction to algorithmic complexity. Before describing the sum-product algorithm, we begin in Section 1.2 with an introduction to its simplest sibling, the variable elimination algorithm. The latter is illustrated over a simple example in order to offer an intuitive understanding of main principles governing the sum-product algorithm. The rest of the chapter is dedicated to the sum-product algorithm. We start in Sections 1.3 and 1.4 with its foundations in two steps. In Section 1.3 we will see how the structure of graphical models is exploited to build a so-called *junction-tree* over which an unnormalized measure factorizes in a compact way for future complexity reduction of an inference. In Section 1.4 we explain how a *belief* is propagated in the junction-tree to obtain quantities of interest using tractable local computations. Sections 1.5 and 1.6 are dedicated to practical aspects of the implementation. Dealing with product of small quantities, one may inevitably encounter underflow issues in practice. We detail in Section 1.5 one of the possible tricks to be applied during the implementation for avoiding such issue. In Section 1.6 some computational shortcuts applied on the graph are given for further enhancement of computational complexity reduction. Finally in Section 1.7 we develop the method over three particular graphs that will be ubiquitous in the thesis and in Section 1.8 we introduce another similar algorithm called *max-product* (or *max-sum*) algorithm for performing other types of inferences, in particular the maximum a posteriori.

Let us firstly mention that main definitions in graph theory, recalled in Appendix A, are assumed to be known. Furthermore, we will adopt throughout the whole manuscript the following simplified notation:

- Let $X = \{X_1, \dots, X_n\}$ be a set of variables and $U \subseteq \{1, \dots, n\}$, we denote by X_U the set $\{X_u\}_{u \in U}$.

- Let X be a variable taking its values in $\mathcal{X}$, we denote by $\sum_X$ the sum over all values taken by X, i.e. $\sum_{x\in\mathcal{X}}$.

Note also that $\mathcal{Z}$ and τ are defined per chapter and they may denote different quantities throughout the thesis.

1.1 Introduction

1.1.1 Bayesian networks

1.1.1.1 Chain rule for Bayesian networks

In this chapter, we restrict our work to discrete random variables. Furthermore, continuous variables introduced in future chapters will be all observed. Notions detailed in this chapter are extendable to networks over (latent) continuous variables and in particular to Gaussian and log-normal distributions with usually more challenging computations.

We consider a finite set of discrete random variables $\{X_1, \ldots, X_n\}$ and a distribution $\mathbb{P}$. Applying the chain rule over their joint distribution, we can write

$$\mathbb{P}(X_1, \ldots, X_n) = \mathbb{P}(X_1)\mathbb{P}(X_2|X_1)\mathbb{P}(X_3|X_1, X_2)\ldots\mathbb{P}(X_n|X_1, \ldots, X_{n-1}). \quad (1.1)$$

For all $u \in \{1, \ldots, n\}$, we denote by $\mathcal{X}_u$ be the set of values taken by X_u. Assuming that, for all $u \in \{1, \ldots, n\}$, $|\mathcal{X}_u| = k$, the number of parameters in $\mathbb{P}$ is $(k-1)+k(k-1)+k^2(k-1)+\ldots+k^{n-1}(k-1) = k^n - 1$ which can be computationally and/or statistically intractable in many statistical problems with an increasing number of variables. Exploiting conditional independencies holding in $\mathbb{P}$ may lead to a drastic drop of complexity to compute Equation (1.1) as we will see in this chapter. We define thereafter the notion of conditional independency and a Bayesian Network (BN).

Definition 1 (conditional independency)**.** *Let X, Y and Z be three disjoint sets of random variables. We say that X and Y are conditionally independent given Z in a distribution $\mathbb{P}$ if and only if $\mathbb{P}(X|Y, Z) = \mathbb{P}(X|Z)$.*

Definition 2 (Bayesian network)**.** *A Bayesian network (BN) is a pair $\mathcal{B} = (G, \mathbb{P})$ where $G = (X = \{X_1, \ldots, X_n\}, \mathcal{E} \subset X \times X)$ is a Directed Acyclic Graph (DAG) and $\mathbb{P}$ is a probability distribution such that $\mathbb{P}(X)$ factorizes according to G. In other words, the joint probability of $X_1, \ldots, X_n$ can be written as:*

$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{u=1}^{n} \mathbb{P}(X_u|X_{\mathrm{pa}(u)}) \quad (1.2)$$

where, for all $u \in \{1, \ldots, n\}$, $X_{\mathrm{pa}(u)}$ is the (possibly empty) set of parents of X_u in G (see Definition 13).

Therefore the joint probability of $X_1, \ldots, X_n$ can be written as a product of Conditional Probability Distributions (CPDs) of the form $\mathbb{P}(X_u|X_{\text{pa}(u)})$. Equation (1.2) is called the chain rule for BNs.

Edges in the DAG of a BN do not hold causal relationships but a set of conditional independence properties called d-separation. Therefore one must not necessarily think of causality when building the structure of a BN but one should ensure that d-separation properties respect our perception of the problem. Whereas d-separation properties are a key for learning the structure of a BN, it is exploited but not questioned for reducing the complexity of an inference over a BN with a known structure. For sustaining the fluidness of the reading, we will not go into a detailed explanation about d-separation but we will see how independencies holding in graphical models are exploited for performing an inference in a BN (see Koller and Friedman, 2009, Section 3.3, for an introduction to d-separation properties).

The following proposition:

Proposition 1. *Let $G = (X = \{X_1, \ldots X_n\}, \mathcal{E})$ be a graph, there exists a topological ordering for $X_1, \ldots, X_n$ in G if and only if G is a DAG*

implies that there exists at least one topological ordering for any DAG. Therefore, imposing that $X_0 = \emptyset$, we can assume that $\text{pa}(1) = \{0\}$ and for all $u \in \{2, \ldots, n\}$, $\text{pa}(u) = \{0\}$ or $\text{pa}(u) \in \{1, \ldots, u-1\}^{|\text{pa}(u)|}$ is in increasing ordering.

Note some particular cases of BNs such as:

- For all $u \in 1, \ldots, n$, $\text{pa}(u) = \{0\} \Rightarrow X_1, \ldots, X_n$ are independent.
- $\text{pa}(1) = \{0\}$ and $\forall u \in \{2, \ldots, n\}, \text{pa}(u) = \{u-1\} \Rightarrow \{X_1, \ldots, X_n\}$ is a Markov chain.
- $v < n$, $\text{pa}(1) = \ldots = \text{pa}(v) = \{0\}$ and for all $u \in \{v+1, \ldots, n\}, \text{pa}(u) = \{u-v, \ldots, u-1\} \Rightarrow \{X_1, \ldots, X_n\}$ is a Markov chain of order v.

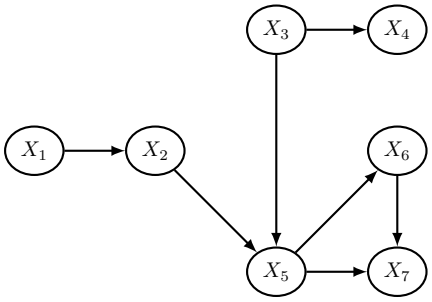
In this whole chapter we denote by $\mathcal{B} = (G = (X = \{X_1, \ldots, X_n\}, \mathcal{E} \subset X \times X), \mathbb{P})$ a BN over a set of n random variables $X = \{X_1, \ldots, X_n\}$ and, for all $u \in \{1, \ldots, n\}$, by $\mathcal{X}_u$ the set of values taken by X_u. Furthermore we will constantly refer to a particular example $\mathcal{B}^{\text{toy}} = (G^{\text{toy}}, \mathbb{P})$ represented Figure 1.1 over a set of seven random variables $X = \{X_1, \ldots, X_7\} \in \{0, 1, 2\}^7$.

1.1.1.2 Evidence and potentials in a Bayesian network

Evidence. We define an evidence in $\mathcal{B}$ to be

$$\text{ev} \stackrel{\text{def}}{=} \cap_{u \in E} \{X_u \in \mathcal{X}_u^* \subset \mathcal{X}_u\} \quad \text{with} \quad E \subset \{1, \ldots, n\}, \tag{1.3}$$

i.e. a subset of values for a subset of variables. For instance $\text{ev} = \{X_3 = 0\}$, $\text{ev} = \{X_7 \neq 0\}$, $\text{ev} = \{X_1 = 0, X_7 \in \{0, 2\}\}$ are some examples of evidences in $\mathcal{B}^{\text{toy}}$. A hard evidence $\{X_E = x_e\}$ where $x_e = (x_u)_{u \in E}$ is a vector of observed values taken by X_E is defined as a subset of variables with a single value assigned.



(a) DAG G^{toy}

$X_1 = 0$	$X_1 = 1$	$X_1 = 2$
0.6	0.3	0.1

(b) $\mathbb{P}(X_1)$

	$X_1 = 0$	$X_1 = 1$	$X_1 = 2$
$X_2 = 0$	0.7	0.6	0.2
$X_2 = 1$	0.2	0.2	0.3
$X_2 = 2$	0.1	0.2	0.5

(c) $\mathbb{P}(X_2|X_1)$

$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
0.6	0.3	0.1

(d) $\mathbb{P}(X_3)$

	$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
$X_4 = 0$	0.7	0.6	0.2
$X_4 = 1$	0.2	0.2	0.3
$X_4 = 2$	0.1	0.2	0.5

(e) $\mathbb{P}(X_4|X_3)$

	$X_2 = 0, X_3 = 0$	$X_2 = 1, X_3 = 0$	$X_2 = 2, X_3 = 0$
$X_5 = 0$	0.9	0.7	0.3
$X_5 = 1$	0.05	0.2	0.5
$X_5 = 2$	0.05	0.1	0.2

	$X_2 = 0, X_3 = 1$	$X_2 = 1, X_3 = 1$	$X_2 = 1, X_3 = 2$
$X_5 = 0$	0.7	0.2	0.1
$X_5 = 1$	0.2	0.5	0.4
$X_5 = 2$	0.1	0.3	0.5

	$X_2 = 0, X_3 = 2$	$X_2 = 1, X_3 = 2$	$X_2 = 2, X_3 = 2$
$X_5 = 0$	0.3	0.1	0.05
$X_5 = 1$	0.5	0.4	0.05
$X_5 = 2$	0.2	0.5	0.9

(f) $\mathbb{P}(X_5|X_2, X_3)$

	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$
$X_6 = 0$	0.7	0.6	0.2
$X_6 = 1$	0.2	0.2	0.3
$X_6 = 2$	0.1	0.2	0.5

(g) $\mathbb{P}(X_6|X_5)$

	$X_5 = 0, X_6 = 0$	$X_5 = 1, X_6 = 0$	$X_5 = 2, X_6 = 0$
$X_7 = 0$	0.9	0.7	0.3
$X_7 = 1$	0.05	0.2	0.5
$X_7 = 2$	0.05	0.1	0.2

	$X_5 = 0, X_6 = 1$	$X_5 = 1, X_6 = 1$	$X_5 = 2, X_6 = 1$
$X_7 = 0$	0.7	0.2	0.1
$X_7 = 1$	0.2	0.5	0.4
$X_7 = 2$	0.1	0.3	0.5

	$X_5 = 0, X_6 = 2$	$X_5 = 1, X_6 = 2$	$X_5 = 2, X_6 = 2$
$X_7 = 0$	0.3	0.1	0.05
$X_7 = 1$	0.5	0.4	0.05
$X_7 = 2$	0.2	0.5	0.9

(h) $\mathbb{P}(X_7|X_5, X_6)$

Figure 1.1: Bayesian network $\mathcal{B}^{\text{toy}} = (G^{\text{toy}}, \mathbb{P})$. (DAG G^{toy} and CPDs).

Potentials. A potential ϕ is a real-valued function over a finite set of variables called the scope of the potential and denoted $\text{Scope}(\phi)$. A multiplicative law, denoted $\times$, $\cdot$ or with no sign, is associated with potentials. Let ϕ_1 and ϕ_2 be two potentials with $\text{Scope}(\phi_1) = \{W, Y\}$ and $\text{Scope}(\phi_2) = \{Y, Z\}$ such that $W \cap Y \cap Z = \emptyset$, $\phi_3 = \phi_1 \times \phi_2$ is defined over $\text{Scope}(\phi_3) = \text{Scope}(\phi_1) \cup \text{Scope}(\phi_2) = \{W, Y, Z\}$ such that, for all $w \in \mathcal{W}$, $y \in \mathcal{Y}$, $z \in \mathcal{S}$, $\phi_3(W = w, Y = y, Z = z) = \phi_1(W = w, Y = y) \times \phi_2(Y = y, Z = z)$ where $\mathcal{W}$, $\mathcal{Y}$, $\mathcal{S}$ are respectively the set of values taken by W, Y and Z. The product of potentials is associated with the following properties:

- commutativity: $\phi_1\phi_2 = \phi_2\phi_1$
- associativity: $(\phi_1\phi_2)\phi_3 = \phi_1(\phi_2\phi_3)$
- Existence of a neutral potential. The neutral potential $\mathbf{1}$ is defined over any scope and takes value one for any value taken by its scope. Let ϕ be a potential, we have $\mathbf{1} \times \phi = \phi$.

Furthermore, let ϕ_1 be a potential over the scope $\text{Scope}(\phi_1) = \{Y, Z\}$ such that $Y \cap Z = \emptyset$, we define the marginalization of Y in ϕ_1 to be $\sum_Z \phi_1 = \phi_2$ such that $\text{Scope}(\phi_2) = \text{Scope}(\phi_1) \setminus Z = Y$ and $\phi_2(Y) = \sum_Z \phi_1(Y, Z)$. We also say that Z is summed out of ϕ_1. Conventional properties imply that $\sum_Y \sum_Z \phi = \sum_Z \sum_Y \phi$ where ϕ is a potential and, denoting ϕ_1 and ϕ_2, two potentials such that $Y \notin \text{Scope}(\phi_1)$, we have $\sum_Y(\phi_1\phi_2) = \phi_1 \sum_Y \phi_2$.

Definition 3 (factor graph)**.** *Let $\phi = \{\phi_1, \ldots, \phi_p\}$ be a set of potentials, the factor graph induced by ϕ is the undirected graph $H_\phi = (X, \mathcal{E} \subseteq X \times X)$ such that $X = \cup_{a=1}^{p} \text{Scope}(\phi_a)$ and $\mathcal{E} = \cup_{a=1}^{p} \{(X_u, X_v); \{X_u, X_v\} \subseteq \text{Scope}(\phi_a)\}$.*

Because each CPD of the form $\mathbb{P}(X_u|X_{\text{pa}(u)})$ in a BN is a potential, we can directly deduce from the following definition of a moral graph:

Definition 4 (moral graph)**.** *Let G be a DAG, the moral graph of G is the undirected graph composed of same nodes and edges plus edges linking each pair of variables having a common child node,*

the following proposition:

Proposition 2. *Let ϕ be the set of CPDs in a BN $\mathcal{B} = (G, \mathbb{P})$, the moral graph of G is the factor graph H_ϕ.*

A graphical representation of the moral graph of G^{toy} is proposed in Figure 1.2.

Entering an evidence. Entering an evidence in a BN consists in replacing its CPDs to a potentials taking value 0 for entries inconsistent with the evidence. Let $\text{ev} = \cap_{u \in E}\{X_u \in \mathcal{X}_u^* \subset \mathcal{X}_u\}$, with $E \subset \{1, \ldots, n\}$, be an evidence in $\mathcal{B}$ and for all $u \in \{1, \ldots, n\} \setminus E$, let $\mathcal{X}_u^* = \mathcal{X}_u$, entering ev in $\mathcal{B}$ consists in replacing each of its CPD of the form $\mathbb{P}(X_u|X_{\text{pa}(u)})$ by the following potential:

$$\phi_u(X_{\{\text{pa}(u),u\}}) = \mathbb{1}_{\{X_v \in \mathcal{X}_v^*, \forall v \in \{\text{pa}(u),u\}\}} \mathbb{P}(X_u|X_{\text{pa}(u)})$$

$X_1 = 0$	$X_1 = 1$	$X_1 = 2$
0.6	0.0	0.0

(a) $\phi_1(X_1)$

	$X_1 = 0$	$X_1 = 1$	$X_1 = 2$
$X_2 = 0$	0.7	0.0	0.0
$X_2 = 1$	0.2	0.0	0.0
$X_2 = 2$	0.1	0.0	0.0

(b) $\phi_2(X_1, X_2)$

$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
0.6	0.3	0.1

(c) $\phi_3(X_3)$

	$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
$X_4 = 0$	0.7	0.6	0.2
$X_4 = 1$	0.2	0.2	0.3
$X_4 = 2$	0.1	0.2	0.5

(d) $\phi_4(X_3, X_4)$

	$X_2 = 0, X_3 = 0$	$X_2 = 1, X_3 = 0$	$X_2 = 2, X_3 = 0$
$X_5 = 0$	0.9	0.7	0.3
$X_5 = 1$	0.0	0.0	0.0
$X_5 = 2$	0.0	0.0	0.0

	$X_2 = 0, X_3 = 1$	$X_2 = 1, X_3 = 1$	$X_2 = 1, X_3 = 2$
$X_5 = 0$	0.7	0.2	0.1
$X_5 = 1$	0.0	0.0	0.0
$X_5 = 2$	0.0	0.0	0.0

	$X_2 = 0, X_3 = 2$	$X_2 = 1, X_3 = 2$	$X_2 = 2, X_3 = 2$
$X_5 = 0$	0.3	0.1	0.05
$X_5 = 1$	0.0	0.0	0.0
$X_5 = 2$	0.0	0.0	0.0

(e) $\phi_5(X_2, X_3, X_5)$

	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$
$X_6 = 0$	0.7	0.0	0.0
$X_6 = 1$	0.2	0.0	0.0
$X_6 = 2$	0.1	0.0	0.0

(f) $\phi_6(X_5, X_6)$

	$X_5 = 0, X_6 = 0$	$X_5 = 1, X_6 = 0$	$X_5 = 2, X_6 = 0$
$X_7 = 0$	0.0	0.0	0.0
$X_7 = 1$	0.05	0.0	0.0
$X_7 = 2$	0.05	0.0	0.0

	$X_5 = 0, X_6 = 1$	$X_5 = 1, X_6 = 1$	$X_5 = 1, X_6 = 2$
$X_7 = 0$	0.0	0.0	0.0
$X_7 = 1$	0.2	0.0	0.0
$X_7 = 2$	0.1	0.0	0.0

	$X_5 = 0, X_6 = 2$	$X_5 = 1, X_6 = 2$	$X_5 = 2, X_6 = 2$
$X_7 = 0$	0.0	0.0	0.0
$X_7 = 1$	0.5	0.0	0.0
$X_7 = 2$	0.2	0.0	0.0

(g) $\phi_7(X_5, X_6, X_7)$

Table 1.1: Potentials in $\mathcal{B}^{\text{toy}}$ obtained after entering ev $= \{X_1 = 0, X_5 = 0, X_7 \in \{1, 2\}\}$.

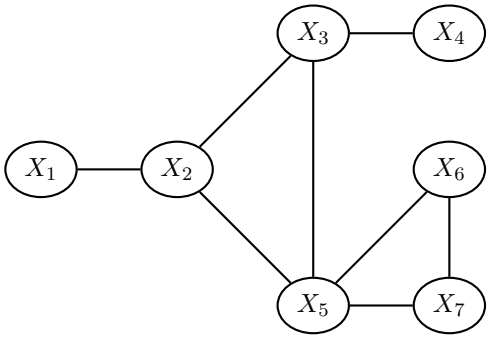


Figure 1.2: Moral graph of G^{toy}.

where $\mathbb{1}_{\{.\}}$ is the indicator function. Table 1.1 lists potentials $\{\phi_u\}_{u=\{1,\dots,7\}}$ obtained from CPDs in $\mathcal{B}^{\text{toy}}$ after entering evidence $\text{ev} = \{X_1 = 0, X_5 = 0, X_7 \in \{1, 2\}\}$.

It follows a natural extension of a hard evidence to uncertain observations, first pointed at by Pearl (1988) under the name of *virtual* and *soft* evidence. Replacing the indicator function by a chosen function ω of a chosen subset of variables in X (in particular X_u but wider *scenarii* can be considered), one can take into account the uncertainty associated with a hard assignment. Let us for instance return to our example $\mathcal{B}^{\text{toy}}$ and suppose that, for a given $u \in \{1, \dots, n\}$, $X_u = o_1$ and $X_u = o_2$ are two observations made by two different observers for X_u. One can define

$$\phi_u(X_{\{\text{pa}(u),u\}}) = \omega(X_u)\mathbb{P}(X_u|X_{\text{pa}(u)})$$

with

$$\omega(X_u) = \mathbb{P}(O_1 = o_1|X_u)\mathbb{P}(O_2 = o_2|X_u),$$

where O_1 and O_2 are two added virtual variables denoting each observer. Note that ω is not a probability measure and in particular it does not necessarily sum to one. One can view a virtual evidence as an extension of the BN with additional variables chosen according to the context. A delicate point of such an evidence is the choice of those variables, their graph parents and added functions.

Recalling the chain rule for a BN $\mathcal{B}$ (Equation 1.2), les ev be an evidence in $\mathcal{B}$, the joint probability of X and ev is given by

$$\mathbb{P}(X, \text{ev}) = \prod_{u=1}^{n} \phi_u(X_{S_u}) \tag{1.4}$$

where for all $u \in \{1, \dots, n\}$, ϕ_u is the potential associated with $\mathbb{P}(X_u|X_{\text{pa}(u)})$ after entering ev in $\mathcal{B}$ and $\text{Scope}(\phi_u) = X_{S_u}$. Note that, so far, we have $X_{S_u} = X_{\{\text{pa}(u),u\}}$ for each u and the factor graph H_ϕ, where $\phi = \{\phi_u\}_{u=1,\dots,n}$ is simply the moral graph of the DAG of the BN but we will see in Section 1.6 a series of computational shortcuts that may lead to reduced scopes.

1.1.2 Algorithmic complexity

A key property of an algorithm is its complexity which evaluates how well it scales both in time and in memory. The complexity in time (respectively in memory)

measures the number of steps taken (respectively of space required) as a function of the number of input data n. As the algorithmic complexity depends of course, among other parameters, on its implementation, it may be multiplied by a constant and the mathematical notation big O is used to express its asymptotical behavior. We say for instance that the complexity of an algorithm is of order $\mathcal{O}(n)$ if it requires a number of operations equal to $C \times n$ where C is a constant. Most algorithms encountered have algorithmic complexities of order $\mathcal{O}(1)$ (constant time), $\mathcal{O}(n)$ (linear time), $\mathcal{O}(\log(n))$ (logarithmic time), $\mathcal{O}(n \log(n))$ (quasi-linear time), $\mathcal{O}(n^k)$ where k is a constant (polynomial time among which quadratic ($k = 2$) and cubic ($k = 3$)), $\mathcal{O}(k^n)$ where k is a contant (exponential time with linear exponent), $\mathcal{O}(k^{\text{poly}(n)})$ where poly(n) is a polynomial in n (exponential time), among others (log-logarithmic, polylogarithmic, factorial, double exponential time, etc.). Algorithms whose complexity is polynomial or exponential become rapidly intractable with an increasing n.

Complexity theory aims at classifying problems in terms of computational cost. A decision problem is a program that accepts an input if it satisfies certain conditions and rejects it otherwise. A decision problem is said to be of class P (polynomial) if there exists a deterministic algorithm that accepts or rejects the input in polynomial time in the size of the input. A decision problem is said to be of class NP (non deterministic polynomial) if a guess to answer the problem can be verified in polynomial time. A decision problem of class NP is divided in two steps. During the first step, the algorithm nondeterministically proposes a guess and during the second step, it verifies it in polynomial time. Clearly $P \subseteq NP$ as a deterministic computation is a special case of a nondeterministic one. A wide open question that found no answer so far in complexity theory is to determine whether $P = NP$ or not.

A reduction is a transformation of a problem into another one. If an algorithm that solves a problem X can be used to solve another problem Y, then Y is no more difficult than X and we say that Y reduces to X. The reduction can be of various complexity. A decision problem is said to be of class NP-hard if any NP problem can be reduced to it with a reduction of polynomial-time complexity. Therefore, NP-hard problems are at least as “hard” as NP ones. NP-complete class is the intersection of NP and NP-hard classes. It contains therefore, somehow, the "hardest" NP problems. Many problems are of class NP-hard and therefore, the search of tractable algorithms for them is a wide area in complexity theory.

We are interested, throughout the whole chapter, in performing an exact inference based on the unnormalized measure:

$$\psi(X) \stackrel{\text{def}}{=} \prod_{a=1}^{p} \phi_a(X_{S_a})$$

where, for all $a \in \{1, \ldots, p\}$, ϕ_a is a potential with $\text{Scope}(\phi_a) = X_{S_a}$ and $X = \cup_{a=1}^{p} X_{S_a}$. A particular example of interest is for instance given by $\psi(X) = \mathbb{P}(X, \text{ev})$ where ev is an evidence in BN $\mathcal{B}$ and $\mathbb{P}(X, \text{ev})$ is given in Equation (1.4). Note that in general, there are not necessarily as many potentials as variables in X, such that $p \leq n$, in particular after applying a series of computational shortcuts but this point is postponed to Section 1.6. A brute force inference is exponential in n and we will see throughout this chapter algorithms based on graphical models for reducing such

time complexity in many encountered problems.

1.2 Variable elimination

The *sum-product variable elimination* algorithm is the first brick of the *sum-product* algorithm also called *belief propagation* algorithm in general graphs and *forward-backward* algorithm in Hidden Markov Models (HMMs). In this introductory section, we explain the principles of the sum-product variable elimination algorithm in a very particular and simple example and reduce our focus on exact inference of posterior probabilities in order to sustain intuition before explaining the sum-product algorithm in the following sections. We are interested in computing the posterior probability of a variable X_u conditional on ev based on Equation (1.4). We consider in this section our simple example $\mathcal{B}^{\text{toy}}$ represented in Figure 1.1 and an evidence $\text{ev} = \cap_{u \in E}\{X_u \in \mathcal{X}_u^* \subset \mathcal{X}_u\}$ where $E \subset \{1, \dots, n\}$. Let $\phi = \{\phi_1, \dots, \phi_7\}$ be the set of potentials associated with CPDs in $\mathcal{B}^{\text{toy}}$ after entering ev.

1.2.1 Principles with an introductory example

Let us for instance consider the query $\mathbb{P}(X_1|\text{ev})$. We first need to compute $\mathbb{P}(X_1, \text{ev})$ by marginalizing X_1 in Equation (1.4):

$$\mathbb{P}(X_1, \text{ev}) = \sum_{X_2} \dots \sum_{X_7} \phi_1(X_1)\phi_2(X_1, X_2)\phi_3(X_3)\phi_4(X_3, X_4)\times \\ \phi_5(X_2, X_3, X_5)\phi_6(X_5, X_6)\phi_7(X_5, X_6, X_7). \quad (1.5)$$

Eliminating a variable in Equation (1.5) consists in summing it out of the expression. Of course the choice of the elimination ordering has no impact on the result but it can have a major impact on the algorithmic complexity. This choice will be discussed later in that section and in a deeper sense in Sections 1.3 and 1.4. Let us randomly choose the elimination ordering $\sigma_1 = (X_7, X_6, X_5, X_4, X_3, X_2)$, we start with X_7. No potential but ϕ_7 contains X_7 in its scope and therefore Equation (1.5) can be rewrite as

$$\mathbb{P}(X_1, \text{ev}) = \sum_{X_2} \dots \sum_{X_6} \phi_1(X_1)\phi_2(X_1, X_2)\phi_3(X_3)\phi_4(X_3, X_4)\times \\ \phi_5(X_2, X_3, X_5)\phi_6(X_5, X_6)\Big(\sum_{X_7} \phi_7(X_5, X_6, X_7)\Big).$$

These independencies are also a direct consequence of d-separation properties between variables. The second variable to eliminate according to σ_1 is X_6. We can again notice that no potential but ϕ_6 and ϕ_7 contains X_6 in its scope and therefore

we get

$$\mathbb{P}(X_1, \text{ev}) = \sum_{X_2} \dots \sum_{X_5} \phi_1(X_1)\phi_2(X_1, X_2)\phi_3(X_3)\phi_4(X_3, X_4)\times \\ \phi_5(X_2, X_3, X_5)\Big\{ \sum_{X_6} \phi_6(X_5, X_6)\Big(\sum_{X_7} \phi_7(X_5, X_6, X_7)\Big)\Big\}.$$

The third variable to be eliminated according to σ_1 is X_5 and no potential but ϕ_5, ϕ_6 and ϕ_7 contains X_5 in its scope, hence we can write

$$\mathbb{P}(X_1, \text{ev}) = \sum_{X_2} \dots \sum_{X_4} \phi(1)\phi_2(X_1, X_2)\phi_3(X_3)\phi_4(X_3, X_4)\times \\ \Big[\sum_{X_5} \phi_5(X_2, X_3, X_5)\Big\{ \sum_{X_6} \phi_6(X_5, X_6)\Big(\sum_{X_7} \phi_7(X_5, X_6, X_7)\Big)\Big\}\Big)\Big].$$

With the same reasoning, no potential but ϕ_4 contains X_4 in its scope and therefore X_4 is simply summed out of ϕ_4:

$$\mathbb{P}(X_1, \text{ev}) = \sum_{X_2} \dots \sum_{X_3} \phi_1(X_1)\phi_2(X_1, X_2)\phi_3(X_3)\Big(\sum_{X_4} \phi_4(X_3, X_4)\Big)\times \\ \Big[\sum_{X_5} \phi_5(X_2, X_3, X_5)\Big\{ \sum_{X_6} \phi_6(X_5, X_6)\Big(\sum_{X_7} \phi_7(X_5, X_6, X_7)\Big)\Big\}\Big)\Big].$$

Pursuing the elimination of the variables according to the chosen ordering σ_1 and using conditional independencies between variables we finally get:

$$\mathbb{P}(X_1, \text{ev}) = \phi_1(X_1)\Big[\sum_{X_2} \phi_2(X_1, X_2)\Big\{ \sum_{X_3} \phi_3(X_3)\Big(\sum_{X_4} \phi_4(X_3, X_4)\Big)\times \\ \Big[\sum_{X_5} \phi_5(X_2, X_3, X_5)\Big\{ \sum_{X_6} \phi_6(X_5, X_6)\Big(\sum_{X_7} \phi_7(X_5, X_6, X_7)\Big)\Big\}\Big]\Big\}\Big].$$

1.2.2 Messages and complexity

Elimination X_7 out of Equation (1.5) consists in locally summing X_7 out of the intermediate potential $\gamma_1 = \phi_7$ and store the result. Note that $\text{Scope}(\gamma_1) = \text{Scope}(\phi_7)$ and γ_1 is indexed by 1 as it is used in the first elimination according to elimination ordering σ_1. We introduce it here to ease future notations but it is unnecessary to store it. The result of the local marginalization $\sum_{X_7} \gamma_1(X_5, X_6, X_7)$ denoted $\delta_{1\to 2}$ is also a potential. We call it a message for reasons explained later. Note that $\text{Scope}(\delta_{1\to 2}) = \text{Scope}(\gamma_1) \setminus X_7 = \{X_5, X_6\}$. The message $\delta_{1\to 2}$ is indexed by $1 \to 2$ as it is the first message computed by a local marginalization respecting ordering σ_1 and it will be used for the second local marginalization. $\delta_{1\to 2}$ will not need to be computed each time another variable is eliminated and therefore it is computed once and for all and stored for later use. Avoiding multiple computations of the same

quantity by storing it and using it when needed is called *dynamic programming*[1] and is the core of variable elimination and the sum-product algorithm.

Using the same reasoning, eliminating X_6 out of Equation (1.5) consists in locally summing it out of the intermediate potential $\gamma_2 = \phi_6\delta_{1\to 2}$. The potential γ_2 is indexed by 2 as it is used in the second elimination and $\text{Scope}(\gamma_2) = \text{Scope}(\phi_6) \cup \text{Scope}(\delta_{1\to 2}) = \{X_5, X_6\}$. The result of the local marginalization $\sum_{X_6} \gamma_2(X_5, X_6)$ is a potential called a message denoted $\delta_{2\to 3}$ as it is the second message computed and it will be later used for the third computation, when eliminating X_5 according to σ_1. Message $\delta_{2\to 3}$ is computed once and for all and stored. Note that $\text{Scope}(\delta_{2\to 3}) = \text{Scope}(\gamma_2) \setminus X_6 = \{X_5\}$.

Similarly X_5 is locally summed out of $\gamma_3 = \phi_{X_5}\delta_{2\to 3}$ to create the message $\delta_{3\to 5}$ indexed by $3 \to 5$ as it will later be used for the fifth local computation when eliminating X_3. We have $\text{Scope}(\gamma_3) = \{X_2, X_3, X_5\}$ and $\text{Scope}(\delta_{3\to 5}) = \text{Scope}(\gamma_3) \setminus X_5 = \{X_2, X_3\}$.

Finally the only quantities needed to compute $\mathbb{P}(X_1, \text{ev})$ and $\mathbb{P}(\text{ev})$ are the set of $m = 5$ messages $\{\delta_{1\to 2}, \delta_{2\to 3}, \delta_{3\to 5}, \delta_{4\to 5}, \delta_{5\to 6}\}$ recursively obtained by local computations:

$$\begin{aligned}
\delta_{1\to 2}(X_5, X_6) &= \sum_{X_7} \gamma_1(X_5, X_6, X_7) = \sum_{X_7} \phi_7(X_5, X_6, X_7);\\
\delta_{2\to 3}(X_5) &= \sum_{X_6} \gamma_2(X_5, X_6) = \sum_{X_6} \phi_6(X_5, X_6)\delta_{1\to 2}(X_5, X_6);\\
\delta_{3\to 5}(X_2, X_3) &= \sum_{X_5} \gamma_3(X_2, X_3, X_5) = \sum_{X_5} \phi_5(X_2, X_3, X_5)\delta_{2\to 3}(X_5)\\
\delta_{4\to 5}(X_3) &= \sum_{X_4} \gamma_4(X_3, X_4) = \sum_{X_4} \phi_4(X_3, X_4);\\
\delta_{5\to 6}(X_2) &= \sum_{X_3} \gamma_5(X_2, X_3) = \sum_{X_3} \phi_{X_3}(X_3)\delta_{3\to 5}(X_2, X_3)\delta_{4\to 5}(X_3) \quad (1.6)
\end{aligned}$$

and finally we have

$$\mathbb{P}(X_1, \text{ev}) = \delta_{6\to\emptyset}(X_1) = \sum_{X_2} \gamma_6(X_1, X_2) = \sum_{X_2} \phi_1(X_1)\phi_2(X_1, X_2)\delta_{5\to 6}(X_2).$$

For all $u \in \{1, \ldots, n\}$, let $\mathcal{X}_u$ be the set of values taken by X_u ($\mathcal{X}_u = \{0, 1, 2\}$ for each variable in our example), for all $i \in \{1, \ldots m\}$, the algorithmic complexity required for each local marginalization $\delta_{i\to\cdot}$ is of order $\mathcal{O}\left(\prod_{X_u \in \text{Scope}(\gamma_i)} |\mathcal{X}_u|\right)$. In our particular example, as $|\mathcal{X}_u|$ is equal for each u, the computational cost of each $\delta_{i\to\cdot}$ is of order $\mathcal{O}\left(|\mathcal{X}_u|^{|\text{Scope}(\gamma_i)|}\right)$. Decomposing indeed, for instance, $\delta_{2\to 3}(X_5) =$

[1] Dynamic programming was developed by Richard Bellman in the 1950s and finds applications in a wide range of domains involving a problem decomposable into several suboptimal problems that recur several times.

$\sum_{X_6} \gamma_2(X_5, X_6)$ and $\delta_{3\to5}(X_2, X_3) = \sum_{X_5} \gamma_3(X_2, X_3, X_5)$ we get

$$\begin{aligned}\delta_{2\to3}(X_5) = \phi_6(X_5, X_6 = 0)\, \delta_{1\to2}(X_5, X_6 = 0) + \\ \phi_6(X_5, X_6 = 1)\, \delta_{1\to2}(X_5, X_6 = 1) + \\ \phi_6(X_5, X_6 = 2)\, \delta_{1\to2}(X_5, X_6 = 2)\end{aligned}$$

which requires 3×3 multiplications and $3 \times (3-1)$ additions and therefore leads to an algorithmic complexity of order $\mathcal{O}(3^2)$ and

$$\begin{aligned}\delta_{3\to5}(X_2, X_3) = \phi_5(X_2, X_3, X_5 = 0)\, \delta_{2\to3}(X_5 = 0) + \\ \phi_5(X_2, X_3, X_5 = 1)\, \delta_{2\to3}(X_5 = 1) + \\ \phi_5(X_2, X_3, X_5 = 2)\, \delta_{2\to3}(X_5 = 2)\end{aligned}$$

which requires 3×3^2 multiplications and $3^2 \times (3-1)$ additions, hence, an algorithmic complexity of order $\mathcal{O}(3^3)$.

This leads to a total complexity of order $\mathcal{O}\left(m \times \arg\max_i \prod_{X_u \in \text{Scope}(\gamma_i)} |\mathcal{X}_u|\right)$ to compute $\mathbb{P}(X_1, \text{ev})$, i.e. $\mathcal{O}\left(m \times \arg\max_i |\mathcal{X}_u|^{|\text{Scope}(\gamma_i)|}\right)$ in the framework of variables of similar cardinality. The power number $|\text{Scope}(\gamma_i)|$ is a direct consequence of the chosen elimination ordering. Making indeed another choice of elimination ordering $\sigma_2 = (X_2, X_7, X_5, X_6, X_4, X_3)$, Equation (1.5) following σ_2 can be rewrite as

$$\mathbb{P}(X_1, \text{ev}) = \phi_1(X_1)\Big[\sum_{X_3} \phi_3(X_3)\Big(\sum_{X_4} \phi_4(X_3, X_4)\Big) \times$$
$$\Big[\sum_{X_6}\Big\{\sum_{X_5} \phi_6(X_5, X_6)\Big(\sum_{X_7} \phi_7(X_5, X_6, X_7)\Big)\Big(\sum_{X_2} \phi_2(X_1, X_2)\phi_5(X_2, X_3, X_5)\Big)\Big\}\Big]\Big]$$

which requires the computation of the following messages:

$$\alpha_{1\to3}(X_1, X_3, X_5) = \sum_{X_2} \beta_1(X_1, X_2, X_3, X_5) = \sum_{X_2} \phi_2(X_1, X_2)\phi_5(X_2, X_3, X_5);$$
$$\alpha_{2\to3}(X_5, X_6) = \sum_{X_7} \beta_2(X_5, X_6, X_7) = \sum_{X_7} \phi_7(X_5, X_6, X_7);$$
$$\alpha_{3\to4}(X_1, X_3, X_6) = \sum_{X_5} \beta_3(X_1, X_3, X_5, X_6) = \sum_{X_5} \phi_6(X_5, X_6)\alpha_{1\to3}(X_1, X_3, X_5)\alpha_{2\to3}(X_5, X_6);$$
$$\alpha_{4\to6}(X_1, X_3) = \sum_{X_6} \beta_4(X_1, X_3, X_6) = \sum_{X_6} \alpha_{3\to4}(X_1, X_3, X_6);$$
$$\alpha_{5\to6}(X_3) = \sum_{X_4} \beta_4(X_3, X_4) = \sum_{X_4} \phi_4(X_3, X_4)$$

and finally we get

$$\mathbb{P}(X_1, \text{ev}) = \alpha_{6\to\emptyset}(X_1) = \sum_{X_3} \beta_5(X_1, X_3) = \sum_{X_3} \phi_1(X_1)\phi_3(X_3)\alpha_{4\to6}(X_1, X_3)\alpha_{5\to6}(X_3).$$

Note that $\arg\max_i |\text{Scope}(\beta_i)| > \arg\max_i |\text{Scope}(\gamma_i)|$ and therefore a higher algorithmic complexity when using elimination ordering σ_2 instead of σ_1. The choice of elimination ordering is one of the key question in the sum-product algorithm and will be detailed in Sections 1.3 and 1.4.

Let us now consider for instance the query $\mathbb{P}(X_3|\text{ev})$ and elimination ordering $\sigma_3 = \{X_7, X_6, X_5, X_1, X_2, X_4\}$, marginalizing X_3 in Equation (1.4) following elimination ordering σ_3 gives

$$\mathbb{P}(X_3, \text{ev}) = \phi_3(X_3)\Big(\sum_{X_4}\phi_4(X_3, X_4)\Big)\Big\{\sum_{X_2}\Big(\sum_{X_1}\phi_1(X_1)\phi_2(X_1, X_2)\Big)\times \Big[\sum_{X_5}\phi_5(X_2, X_3, X_5)\Big\{\sum_{X_6}\phi_6(X_5, X_6)\Big(\sum_{X_7}\phi_7(X_5, X_6, X_7)\Big)\Big\}\Big]\Big\}.$$

Note that the first three messages and the last message computed while eliminating X_7, X_6, X_5 and X_4 to infer $\mathbb{P}(X_3, \text{ev})$ are precisely $\delta_{1\to 2}(X_5, X_6)$, $\delta_{2\to 3}(X_5)$, $\delta_{3\to 5}(X_2, X_3)$ and $\delta_{4\to 5}(X_3)$ previously computed and stored when computing $\mathbb{P}(X_1, \text{ev})$ following elimination ordering σ_1. Dynamic programming and storage of messages $\delta_{\cdot\to\cdot}$ allow for a complexity reduction not only for computing $\mathbb{P}(X_1, \text{ev})$ but also for computing posterior probabilities of each other variables in $\mathcal{B}^{\text{toy}}$.

Let $\phi = \{\phi_a\}_{a=1,\dots,p}$ be a set of potentials, $X = \cup_{a=1}^{p}\text{Scope}(\phi_a)$, $Y \subseteq X$ and σ be an elimination ordering over Y, the sum-product variable elimination algorithm over $\sum_Y \prod_{a=1}^{p}\phi_a$ denoted $\text{VE}(\phi, Y, \sigma)$ is developed in Algorithm 1.

Algorithm 1: Sum-product variable elimination algorithm

ϕ : set of potentials;
Y : set of N variables to be eliminated;
$\sigma = (Y_1, \dots, Y_N)$: elimination ordering over Y;
for i in $1, \dots, N$ **do**
  $\widetilde{\phi} \leftarrow \{\phi_a \in \phi \text{ such that } Y_i \in \text{Scope}(\phi_a)\}$;
  $\gamma \leftarrow \prod_{\phi_a \in \widetilde{\phi}}\phi_a$;
  $\delta \leftarrow \sum_{Y_i}\gamma$;
  $\phi \leftarrow \phi \cup \delta \setminus \widetilde{\phi}$: update ϕ;
end
return $\prod_{\phi_a\in\phi}\phi_a$

We now generalize notions seen in this section with the sum-product algorithm.

1.3 From a Markov network to a junction-tree

In the previous section, we developed the variable elimination algorithm over a simple example in order to introduce notation and quantities needed for the sum-product algorithm also called belief propagation in general graphs and forward-backward algorithm in Hidden Markov Models (HMMs) (Rabiner, 1989). The sum-product

algorithm is a class of *message-passing* algorithms and exploits conditional independence assumptions for reducing the complexity of an inference in probabilistic graphical models with latent variables using local computations and dynamic programming (Lauritzen, 1996). It was first introduced by Pearl (Pearl, 1982, 1986, 1988), originally developed for trees and polytrees and later extended to general graphs (Lauritzen and Spiegelhalter, 1988). The treewidth of a graph is a key notion for bounding the algorithmic complexity of an exact inference as we will see in this section and the next one. Arnborg et al. (1987) proved that computing the treewidth of a graph is a NP-hard problem, however, heuristics exist for providing an upper bond of the treewidth in $\mathcal{O}(n \times \ell)$, where n (respectively ℓ) is the number of nodes (respectively vertices), with good results in practice (see Peyrard et al., 2019, for a synthetic review). The scope of this thesis is restricted to graphs leading to tractable inference, i.e. graphs of limited treewidth. Approximate inference for graphs of large treewidth gathering sampling based methods and variational methods fall outside the scope of this thesis. We recommend for instance the empirical review (Murphy et al., 2013) for readers interested in the most encountered approximate method called the loopy belief propagation algorithm and (Wainwright and Jordan, 2008) for broader types of variational methods. Among several other references, the rest of this chapter was deeply inspired by Pearl (1986); Lauritzen and Spiegelhalter (1988); Shafer and Shenoy (1990); Lauritzen (1992, 1996); Cowell et al. (1999); Jensen and Nielsen (2007); Koller and Friedman (2009).

For the rest of the chapter, in order to lighten notation, a non-oriented (respectively oriented) edge between two variables Y and Z will be denoted $Y - Z$ (respectively $Y \to Z$).

1.3.1 Factorization over a Markov network

Our willing is to perform an inference based on the equation introduced in Section 1.1.2 and recalled thereafter

$$\psi(X) \stackrel{\text{def}}{=} \prod_{a=1}^{p} \phi_a(X_{S_a}) \tag{1.7}$$

where each ϕ_a is a potential over the scope X_{S_a} and $X = \cup_{a=1}^{p} X_{S_a}$. A brute force computation of $\sum_X \psi(X)$ is exponential in the number of variables in X. However, one can exploit the local and global Markov property in the factor graph H_ϕ where $\phi = \{\phi_a\}_{a=1,\dots,p}$ (see Definition 3 of a factor graph) for a complexity reduction. Let us recall these properties:

- Local Markov property: A variable is independent of all other variables conditional on its neighbors. For any $X_u \in X$, $(X_u \perp\!\!\!\perp \{X \setminus X_{\{u,\text{nb}(u)\}}\} | X_{\text{nb}(u)})$ where $\text{nb}(u)$ is the set of labels of neighbors of X_u (see Definition 13).

- Global Markov property: Let X_U, X_V and X_S be three disjoint subsets of X such that X_S separates X_U and X_V, i.e. all path from any variable in X_U to any variable in X_V intersect X_S, then $(X_U \perp\!\!\!\perp X_V | X_S)$.

Let us first notice that $\psi(X)/\sum_X \psi(X)$ factorizes over H_ϕ (see Definition 5).

Definition 5 (Factorization over a Markov network)**.** *We say that a distribution $\mathcal{P}$ factorizes over a Markov network $H = (X, \mathcal{E})$ if there exists a set of potentials $\{\Phi_1(X_{S_1}), \ldots, \Phi_p(X_{S_p})\}$ such that $X = \cup_{a=1}^p X_{S_a}$,*

$$\mathcal{P}(X) = \frac{1}{\mathcal{Z}} \prod_{a=1}^{p} \Phi_a(X_{S_a}) \quad \textit{where} \quad \mathcal{Z} = \sum_X \prod_{a=1}^{p} \Phi_a(X_{S_a}) \tag{1.8}$$

and, for all $a \in \{1, \ldots, p\}$, X_{S_a} is a clique of H, i.e. a complete subgraph of H (see Definitions 14 and 15). The quantity $\mathcal{Z}$ is called the partition function of $\mathcal{P}(X)$.

Furthermore, for reasons developed in Section 1.4, if H is chordal, i.e. any of its cycle of length strictly greater than three has a chord (see Definition 28), the algorithmic complexity to compute $\mathcal{Z}$ or the marginal of any variable in Equation (1.8) is of the order of the sum of all maximal clique configurations. Given an unnormalized measure $\psi(X)$ as written in Equation (1.7), the first step of the sum-product algorithm consists in creating m subsets of X denoted $C_1, \ldots, C_m$ such that $H_{\{\Phi_1(C_1),\ldots,\Phi_m(C_m)\}}$ is chordal, $\psi(X)$ factorizes over $H_{\{\Phi_1(C_1),\ldots,\Phi_m(C_m)\}}$ and $\sum_{i=1}^m \prod_{X_u \in C_i} |\mathcal{X}_u|$ is minimal.

Returning to the variable elimination algorithm detailed in Algorithm 1, we introduce thereafter the definition of a graph induced by a set of potentials and an elimination ordering.

Definition 6 (graph induced by a set of potentials and an elimination ordering)**.** *Let $\phi = \{\phi_1, \ldots, \phi_p\}$ be a set of potentials, $X = \cup_{a=1}^p \text{Scope}(\phi_a)$ and σ an elimination ordering over X, the graph induced by ϕ and σ is the undirected graph $H_{\phi,\sigma} = H_\gamma$ where γ is the set of intermediate potentials created during variable elimination* $\text{VE}(\phi, X, \sigma)$.

Theorem 1. *Any graph induced by a set of potentials $\{\phi_1, \ldots, \phi_p\}$ and an elimination ordering over $\cup_{a=1}^p \text{Scope}(\phi_a)$ is chordal.*

Proof. Let $H_{\phi,\sigma}$ be the graph induced by $\phi = \{\phi_1, \ldots, \phi_p\}$ and an elimination ordering σ over $X = \cup_{a=1}^p \text{Scope}(\phi_a)$. Without loss of generality, let $X_1 - X_2 - \ldots - X_k - X_1$ be a cycle in $H_{\phi,\sigma}$ and let X_1 be the first variable to be eliminated. Because edges $X_1 - X_2$ and $X_1 - X_k$ exist, X_2 and X_k belong to the scope of an intermediate potential γ_i created during variable elimination $\text{VE}(\phi, X, \sigma)$ and consequently, they are linked by an edge in $H_{\phi,\sigma}$. □

If such an edge does not exist in the original factor graph H_ϕ, we call it a fill-in edge. A factor graph H_ϕ is chordal if and only if there exists an elimination ordering σ over X such that $H_\phi = H_{\phi,\sigma}$, in other words, such that running $\text{VE}(\phi, X, \sigma)$ does not add any fill-in edge in H_ϕ. If H_ϕ is chordal and $H_\phi = H_{\phi,\sigma}$ then σ is said to be perfect for H_ϕ.

Definition 7 (triangulation)**.** *Triangulating a non-chordal undirected graph consists in adding fill-in edge(s) to render it chordal.*

Any graph $H_{\phi,\sigma}$ is chordal and contains at least all edges in H_ϕ, therefore, if H_ϕ is non-chordal, building $H_{\phi,\sigma}$ is equivalent to triangulating H_ϕ. Note that the treewidth of a triangulated graph is the size (in terms of number of variables) of its largest clique minus one.

Definition 8 (junction-tree)**.** *Let $H = (X, \mathcal{E})$ be an undirected graph, the undirected graph $J = \big(C = \{C_1, \ldots, C_m\}, \mathcal{F}\big)$ such that for all $i \in \{1, \ldots, m\}$, $C_i \subseteq X$ is a junction-tree (JT) for H if and only if it satisfies the following properties:*

- *J is a tree (tree property).*
- *Whenever a variable (or a set of variables) $X_u \subseteq C_i \cap C_k$, then $X_u \subseteq C_j$ for all C_j in the (unique) path between C_i and C_k (Running intersection property).*
- *For any $(X_u, X_v) \in \mathcal{E}$, $\exists i \in \{1, \ldots, m\}$ such that $\{X_u, X_v\} \subseteq C_i$ (Covering property).*

The *treewidth* of a JT $J = \big(C = \{C_1, \ldots, C_m\}, \mathcal{F}\big)$ is $\tau = \max_{i=1,\ldots,m} |C_i| - 1$. As intuitively understood and proven in Section 1.4, the smallest the treewidth of a triangulated factor graph, hence an associated JT, the smallest the complexity for performing an inference based on a distribution that factorizes over it. Note that for any graph $H = (X, \mathcal{E})$, $J = (X, X \times X)$ composed of a unique large clique is a (bad choice of) JT for H and therefore, there always exists at least one JT for any undirected graph. For some classes of undirected graphs, treewidth leading to tractable exact inference can not be obtained and approximate inference (outside the scope of this thesis) is usually needed.

Proposition 3. *Let $\phi = \{\phi_1, \ldots, \phi_p\}$ be a set of potentials and for all $a \in \{1, \ldots, p\}$, let $X_{S_a} = \text{Scope}(\phi_a)$. Let $X = \cup_{a=1}^{p} X_{S_a}$ and σ be an elimination ordering over X. Then $\text{VE}(\phi, X, \sigma)$ defines a JT for H_ϕ.*

Before proving Proposition 3, let us first introduce some conventional notation and definitions associated with a JT defined by $\text{VE}(\phi, X, \sigma)$. From the definition of a graph $H_{\phi,\sigma}$ induced by ϕ and σ (see Definition 6), let $\gamma = (\gamma_1, \ldots, \gamma_m)$ be the set of intermediate potentials generated during $\text{VE}(\phi, X, \sigma)$ in that order, for all $i \in \{1, \ldots, m\}$, $\text{Scope}(\gamma_i)$ is a clique in $H_{\phi,\sigma}$ denoted C_i. We define $C = \{C_1, \ldots, C_m\}$ to be the set of those cliques. The clique $C_m = \text{Scope}(\gamma_m)$ is said to be the root clique and it contains the last variable eliminated during $\text{VE}(\phi, X, \sigma)$. Although a JT is undirected, the proof lays on directed edges according to the sequential creation of intermediate potentials. When an intermediate potential γ_i is used to create the message $\delta_{i \to j}$ which will be used to build γ_j, we define $j = \text{to}(i)$ and any clique on the path $C_{\text{to}(i)} - \ldots - C_m$ is said te be upstream C_i.

Proof of Proposition 3. The proof is divided in three parts for considering each property of a JT.

1. Tree property: For all $i \in \{1, \ldots, m-1\}$, $\text{to}(i)$ is unique as the message $\delta_{i \to j}$ will be used once and only once to compute γ_j over the scope C_j. Consequently, the resulting graph is a tree.

2. Running intersection property: Let X_u be a variable appearing in two different cliques C_i and C_k such that $k > i$. Each variable is summed out once and only once during $\mathrm{VE}(\phi, X, \sigma)$. Without loss of generality, assume that X_u is eliminated from C_k, i.e. it is summed out of γ_k while computing $\delta_{k\to\cdot}$. Therefore, X_u does not appear in the scope of any intermediate potential created after γ_i during $\mathrm{VE}(\phi, X, \sigma)$ and in particular, it does not appear in any clique upstream C_k. Secondly, because X_u is not summed out of γ_i it belongs to $\mathrm{Scope}(\delta_{i\to \mathrm{to}(i)})$ and $\mathrm{Scope}(\gamma_{\mathrm{to}(i)}) = C_{\mathrm{to}(i)}$. By induction, it belongs to all cliques on the path $C_i - \ldots - C_k$.

3. Covering property: By definition of the graph $H_{\phi,\sigma}$ induced by ϕ and σ, $H_{\phi,\sigma}$ contains at least all edges in H_ϕ.

□

Given a set of potentials $\phi = \{\phi_a\}_{a=1,\ldots,p}$, the choice of an elimination ordering σ over $X = \cup_{a=1}^{p} \mathrm{Scope}(\phi_a)$ leading to a JT defined by $\mathrm{VE}(\phi, X, \sigma)$ is crucial for complexity reduction of a future inference. This point is developed in the next section.

1.3.2 Search for an elimination ordering

Search for an elimination ordering in chordal graphs. We have seen in the previous section that if H_ϕ is chordal, there exists and elimination ordering σ such that $H_\phi = H_{\phi,\sigma}$. Such an elimination ordering is said to be perfect for H_ϕ because no fill-in edge is added leading to a minimal complexity to compute Equation (1.8). It follows directly from the definition of a chordal graph that a subsequent removal of simplicial nodes in a chordal graph leads to a perfect elimination ordering. For instance, note that the moral graph of the DAG of our particular example $\mathcal{B}^{\mathrm{toy}}$ represented in Figure 1.2 is chordal and $\sigma_1 = \{X_7, X_6, X_5, X_4, X_3, X_2, X_1\}$ chosen when presenting the variable elimination algorithm in Section 1.2 over $\mathcal{B}^{\mathrm{toy}}$ and an evidence $\mathrm{ev} = \cap_{u\in E}\{X_u \in \mathcal{X}_u^* \subset \mathcal{X}_u\}$ was obtained by a subsequent removal of simplicial nodes.

Alternatively running, over a chordal graph, the *lexicographic breadth-first search* or *Lex-BFS* (Rose et al., 1976) revisited by Tarjan and Yannakakis (1984) under the name of *maximum cardinality search* algorithm (see Algorithm 2) returns the reverse of a perfect elimination ordering in linear time complexity in the number of variables.

Let us return for instance to our example $\mathcal{B}^{\mathrm{toy}}$ with evidence $\mathrm{ev} = \cap_{u\in E}\{X_u \in \mathcal{X}_u^* \subset \mathcal{X}_u\}$. Let ϕ be the set of potentials associated with CPDs in $\mathcal{B}^{\mathrm{toy}}$ after entering ev, the factor graph H_ϕ (i.e. the moral graph of G^{toy}) represented Figure 1.2 is chordal. All nodes are initialized with weight 0. The first step consists in randomly choose one variable, say X_3. Neighbors of X_3, ie. X_2, X_4 and X_5, are given weight 1. X_3 is removed and a node among $\{X_2, X_4, X_5\}$ is sampled, for instance X_2. Weights of X_1, X_5 are incremented by one which leads to two for X_5 and one for X_1 and X_2 is removed. X_5 being the heaviest variable is chosen, weights of X_6 and X_7 become one and X_5 is removed. All remaining variables X_1, X_4, X_6 and X_7 have same weight equal one. Let us assume that X_6 is randomly sampled. The weight of X_7 becomes

Algorithm 2: Maximum cardinality search

X : set of n nodes;
$\mathcal{E}$: set of edges;
β : empty vector of size n;
initialization: Initialize all nodes in X with weight 0;
for i in $1, \ldots, n$ **do**
    $u \leftarrow$ sample one node with max weight in X;
    $\beta_i \leftarrow u$;
    Increment weight of neighbors of u by 1;
    Remove u;
end
return β

two, X_6 is removed and X_7 becomes the heaviest variable hence, it is chosen. X_7 has no remaining neighbor, it is simply removed. The remaining variables X_1 and X_4 have same weight and are subsequently chosen in removed, for instance in order (X_4, X_1). In the end, the algorithm returns $\beta = (X_3, X_2, X_5, X_6, X_7, X_4, X_1)$ whose ordering is reversed to get $\sigma = (X_1, X_4, X_7, X_6, X_5, X_2, X_3)$. Finally we notice that following ordering σ in H_ϕ leads to the subsequent removal of simplicial nodes, hence σ is a perfect elimination ordering.

Search for elimination orderings in non-chordal graphs. If the factor graph H_ϕ is non-chordal, the triangulation step is the key one for bounding computational complexity. Finding a best elimination ordering σ, hence best fill-in edges to add when triangulating H_ϕ to render it chordal leading to minimal complexity for an inference over $H_{\phi,\sigma}$ is an NP-complete problem (Yannakakis, 1981), so is the determination of the expected bounded treewidth (Arnborg et al., 1987). Cooper (1990) proved that a probabilistic inference in a Bayesian network is NP-hard in general. However, without knowing the resulting complexity and with no tractable algorithm to reach minimal complexity, heuristics exist with good results in practice. Main current heuristics named min-fill, min-weight, weighted min-fill and min-neighbors evaluate the cost of a node if eliminated and differ by the cost function used. They have been surveyed and evaluated by Kjærulff (1990) and show good performances empirically. Intuitively, as the goal is to minimize the sum of the total state space sizes of the cliques in the resulting triangulated graph, the idea is to choose fill-in edges that minimize that quantity both in terms of number of component variables in cliques and their total state spaces. Defining the weight of a node to be the cardinality of its state space, the cost of a node in a given graph is defined as:

- min-fill: the number of edges that must be added if the node is eliminated.
- min-neighbors: the number of its neighbors.
- min-weight: the product of the weight of its neighbors.

- weighted-min-fill: the sum of weights of the edges that must be added if the node is eliminated where the weight of an edge is defined as the product of weights of the nodes it links.

None of these heuristics are better than one another and their performances depend on the context. Min-weight and weighted-min-fill tend to perform better empirically in graphs containing variables with large weight differences. The greedy triangulation algorithm developed in Algorithm 3 returns an elimination ordering using one of these heuristics as cost function for triangulation.

Algorithm 3: Greedy triangulation

X : set of n nodes;
$\mathcal{E}$: set of edges;
c : cost function;
σ : empty vector of size n;
for i in $1, \dots, n$ **do**
  $u \leftarrow$ sample one node in X that minimizes $c(X, \mathcal{E})$;
  $\sigma_i \leftarrow u$;
  Add edges if missing between all neighbors of u;
  Remove u;
end
return σ

In its stochastic version, one node among a set of nodes of minimal cost is stochastically sampled at the selection step.

1.4 Message passing or belief propagation

In a second phase called belief propagation, messages, such as those denoted $\delta_{\cdot\to\cdot}$ in Section 1.2, are passed between cliques of the JT and used for computing the marginal distribution of a clique or a separator. In this section we begin with important definitions and properties before detailing the practical implementation of messages. In order to lighten notation, an edge $C_i - C_j$ under a sum, product or union symbol will be denoted $i - j$.

Let us consider $X = \{X_1, \dots, X_n\}$, a set of variables and $\phi = \{\phi_1, \dots, \phi_p\}$, a set of potentials such that, for all $a \in \{1, \dots, p\}$, $\text{Scope}(\phi_a) = X_{S_a}$ and $X = \cup_{a=1}^{p} X_{S_a}$. We return to the unnormalized measure written in Equation (1.7) and recalled thereafter:

$$\psi(X) \stackrel{\text{def}}{=} \prod_{a=1}^{p} \phi_a(X_{S_a})$$

and, for all $V \subset \{1, \dots, n\}$, $\psi(X_V) = \sum_{X \setminus X_V} \prod_{a=1}^{p} \phi_a(X_{S_a})$ where $X_V = \{X_u\}_{u \in V}$.

1.4.1 Definition of messages

Let $J = (C = (C_1, \ldots, C_m), \mathcal{F})$ be a JT defined by $\mathrm{VE}(\phi, X, \sigma)$ where σ is an elimination ordering over X (see Proposition 3) , we recall that, for all $i \in \{1, \ldots, m\}$, C_i is the scope of the intermediate potential γ_i created during $\mathrm{VE}(\phi, X, \sigma)$ and that C_m, called the root clique, contains the last variable eliminated. The covering property of J (see Definition 8) implies that, for all $a \in \{1, \ldots, p\}$, there exists $i \in \{1, \ldots, m\}$ such that $X_{S_a} \subseteq C_i$. One such clique C_i is chosen for each X_{S_a} and we say that the potential ϕ_a is injected in C_i. We define C_i^* to be the set of potentials injected in the clique C_i. For any edge $C_i - C_j$ we define the separator $S_{i,j} = S_{j,i} = C_i \cap C_j$. Furthermore we define the *upstream* set $U_{i\to j}$ (respectively $U^*_{i\to j}$) to be the union of all cliques (respectively potentials injected in cliques) in the connected component containing C_i in trees obtained from removing the edge $C_i - C_j$ from J.

Definition 9 (message)**.** *For each edge $C_i \to C_j$ we define the message $\delta_{i\to j}(S_{i,j})$ to be*

$$\delta_{i\to j}(S_{i,j}) \stackrel{\text{def}}{=} \sum_{U_{i\to j} \setminus S_{i,j}} \prod_{\phi_a \in U^*_{i\to j}} \phi_a(X_{S_a}).$$

We say that the clique C_i sends the message $\delta_{i\to j}(S_{i,j})$ to C_j and the clique C_j receives it from C_i.

Note that two messages per undirected edge $C_i - C_j$ are defined: $\delta_{i\to j}(S_{i,j})$ and $\delta_{j\to i}(S_{i,j})$.

1.4.2 Implementation and propagation

For each clique C_i, we define the potential of C_i to be

$$\Phi_i(C_i) \stackrel{\text{def}}{=} \begin{cases} \prod_{\phi_a \in C_i^*} \phi_a(X_{S_a}) & \text{if } C_i^* \neq \emptyset \\ 1 & \text{otherwise} \end{cases}. \tag{1.9}$$

Note that, because each potential ϕ_a is injected in one and only one clique of J, we have $\prod_{i=1}^m \Phi_i(C_i) = \prod_{a=1}^p \phi_a(X_{S_a})$.

A recursive implementation of messages is rendered possible by the following corollary:

Corollary 1. *For all $i, j \in \{1, \ldots, m\}$ such that $C_i \to C_j$,*

$$\delta_{i\to j}(S_{i,j}) = \sum_{C_i \setminus S_{i,j}} \Phi_i(C_i) \prod_{k \neq j, k - i} \delta_{k\to i}(S_{k,i}).$$

Proof. Starting from the definition of messages and clique potentials, the proof is simply done by induction. □

We now detail the practical implementation of messages using a topological ordering of cliques. A root clique C_m is randomly chosen and, with respect to running intersection property of J, cliques are assigned indexes in a topological ordering such

that for $i \in \{1, \dots, m-1\}$, $\text{to}(i) > i$. The first step, called *inward* or *forward pass*, consists in recursively computing, for $i = 1, \dots, m-1$, and for all j such that $C_i \to C_j$, the message $\delta_{i \to j}(S_{i,j})$. The second step called *outward* or *backward pass* consists in computing, for $i = m-1, \dots, 1$, and for all j such that $C_j \to C_i$, the message $\delta_{j \to i}(S_{i,j})$. Note that $\delta_{i \to j}(S_{i,j})$ can actually be computed as soon as the clique C_i is *ready* to send it, i.e. as soon as C_i received all messages $\delta_{k \to i}(S_{k,i})$ with $k \neq j$ but in practice, using a topological ordering of cliques allows for a simpler implementation as detailed in Algorithm 4 where $\text{nb}(i)$ denotes the indexes of neighbors of C_i and $\alpha = (\alpha_1, \dots, \alpha_p)$, the vector of assignments of initial potentials to cliques, is such that for all $a \in \{1, \dots, p\}$, $\phi_a \in C^*_{\alpha_a}$. For readability, a message computed during the inward (respectively outward) pass is denoted $\delta_\cdot$ (respectively $\beta_\cdot$).

Algorithm 4: Sum-product algorithm

ϕ: set of potentials (size p);
J: JT s.t. $C_1, \dots, C_m$ are in topological ordering and C_m is a chosen root;
α: vector of assignments of initial potentials to cliques (size p);
Φ, δ, β: empty lists of size m;

1: **procedure** BUILD CLIQUE POTENTIALS(ϕ, α);
  for i in $1, \dots, m$ **do**
    $\Phi_i \leftarrow \prod_{a, \alpha_a = i} \phi_a$;
  end
  return Φ;

procedure INWARD PASS(J, Φ);
  for i in $1, \dots, m$ **do**
    $\gamma \leftarrow \Phi_i \times \prod_{j \in \text{nb}(i) \setminus \text{to}(i)} \delta_j$;
    $\delta_i = \sum_{C_i \setminus S_{i,j}} \gamma$;
  end
  return δ;

procedure OUTWARD PASS(J, Φ, δ);
  Initialization: $\beta_m = 1$;
  for i in $m, \dots, 1$ **do**
    for $j \in \text{nb}(i) \setminus \text{to}(i)$ **do**
      $\gamma \leftarrow \Phi_i \times \beta_i \times \prod_{k \in \text{nb}(i) \setminus \{j, \text{to}(i)\}} \delta_k$;
      $\beta_j = \sum_{C_i \setminus S_{i,j}} \gamma$;
    end
  end
  return β;

return Φ, δ, β

For all $u \in \{1, \dots, n\}$, let $\mathcal{X}_u$ be the set of values taken by X_u, each (inward and outward) pass requires a time complexity of order $\mathcal{O}\left(\sum_{i=1}^{m} \prod_{X_u \in C_i} |\mathcal{X}_u|\right)$. Furthermore, we will see in Section 1.4.3 that clique potentials and messages must be stored for performing certain inference types leading to a memory complexity of order

$\mathcal{O}\big(\sum_{i=1}^{m}\prod_{X_u\in C_i}|\mathcal{X}_u| + 2\times\sum_{i=1}^{m}\prod_{X_u\in S_{i,j}}|\mathcal{X}_u|\big) = \mathcal{O}\left(\sum_{i=1}^{m}\prod_{X_u\in C_i}|\mathcal{X}_u|\right)$.

1.4.3 Exact inference

In this section we introduce theorems for computing marginal distributions based on the following lemma:

Lemma 1. *We have*

- $(U_{i\to j}\setminus S_{i,j})\sqcup(U_{j\to i}\setminus S_{i,j}) = X\setminus S_{i,j}$
- $U_{i\to j}\cap U_{j\to i}\subset S_{i,j}$
- $\sqcup_{i,i-j}U_{i\to j}\setminus C_j = X\setminus C_j$
- *For all* $i'\neq i$, $U_{i'\to j}\cap U_{i\to j}\subset C_j$

Proof. As $X=\cup_{i=1}^{m}C_i$, the proof of the first and third items is straightforward. The second and fourth items are a direct consequence of the running intersection property of a JT. □

Based on Lemma 1 we can deduce the following theorems:

Theorem 2 (marginal of a separator)**.** *For each separator* $S_{i,j}$ *we have*

$$\psi(S_{i,j}) = \delta_{i\to j}(S_{i,j})\delta_{j\to i}(S_{i,j})$$

Proof. Using the first item of Lemma 1 we have

$$\sum_{X\setminus S_{i,j}}\prod_{a=1}^{p}\phi_a(X_{S_a}) = \sum_{U_{i\to j}\setminus S_{i,j}}\sum_{U_{j\to i}\setminus S_{i,j}}\prod_{a=1}^{p}\phi_a(X_{S_a})$$

The second item of Lemma 1 tells us that the only variables in common in the two sums are in $S_{i,j}$ and therefore we can write

$$\sum_{X\setminus S_{i,j}}\prod_{a=1}^{p}\phi_a(X_{S_a}) = \sum_{U_{i\to j}\setminus S_{i,j}}\prod_{\phi_a\in U^*_{i\to j}}\phi_a(X_{S_a})\sum_{U_{j\to i}\setminus S_{i,j}}\prod_{\phi_a\in U^*_{j\to i}}\phi_a(X_{S_a})$$

which concludes the proof. □

Theorem 3 (marginal of a clique)**.** *For each clique* C_i *we have*

$$\psi(C_i) = \Phi_i(C_i)\prod_{j,i-j}\delta_{j\to i}(S_{i,j})$$

Proof. First of all we have

$$\sum_{X\setminus C_i}\prod_{a=1}^{p}\phi_a(X_{S_a}) = \prod_{\phi_a\in C_i^*}\phi_a(X_{S_a})\sum_{X\setminus C_i}\prod_{\phi_a\setminus C_i^*}\phi_a(X_{S_a}).$$

Using the third and fourth items of Lemma 1 we can write

$$\sum_{X\setminus C_i}\prod_{a=1}^{p}\phi_a(X_{S_a}) = \prod_{\phi_a\in C_i^*}\phi_a(X_{S_a})\prod_{j,j-i}\sum_{U_{j\to i}\setminus C_i}\prod_{\phi_a\in U_{j\to i}^*}\phi_a(X_{S_a}),$$

and finally, the running intersection property implies that $U_{j\to i}\setminus C_i = U_{j\to i}\setminus S_{i,j}$, hence we get

$$\sum_{X\setminus C_i}\prod_{a=1}^{p}\phi_a(X_{S_a}) = \prod_{\phi_a\in C_i^*}\phi_a(X_{S_a})\prod_{j,j-i}\sum_{U_{j\to i}\setminus S_{i,j}}\prod_{\phi_a\in U_{j\to i}^*}\phi_a(X_{S_a})$$

which concludes the proof. □

Remark 1. Theorem 3 can be generalized to any subset of X as combining adjacent cliques in a JT produces another JT. Indeed, let $C_r - \ldots - C_s$ be a path in a JT and X_V be a subset of X such that $X_V \subset \{C_r, \ldots, C_s\}$, we have

$$\psi(X_V) = \sum_{\{C_r,\ldots,C_s\}\setminus X_V}\prod_{i=r}^{s}\Phi_i(C_i)\prod_{\substack{j,k,j-k\\ j\notin\{r,\ldots,s\}\\ k\in\{r,\ldots,s\}}}\sum_{C_j\setminus S_{j,k}}\delta_{j\to k}(S_{j,k})$$

Note however that the resulting complexity is exponential in the number of variables in selected cliques.

Remark 2. The partition function $\mathcal{Z} = \sum_X \psi(X)$ can be computed over any separator or any clique. Indeed, for any separator $S_{i,j}$ and any clique C_i we have $\mathcal{Z} = \sum_{S_{i,j}}\psi(S_{i,j}) = \sum_{C_i}\psi(C_i)$. In particular, if no other inference is needed, one can choose C_m and avoid the outward pass of the sum-product algorithm.

Let us finally conclude this section with a sampling method based on the following corollary:

Corollary 2. *The function $\psi(X)$ can fully be described as an heterogeneous Markov tree over J.*

Proof. Choosing any clique C_i as a starting click we have $\mathbb{P}(C_i) \propto \psi(C_i)$. Then recursively, for all $j \in \{1,\ldots,m\}\setminus i$ such that $C_i - C_j$, we have $\mathbb{P}(C_j|C_i) = \mathbb{P}(C_j|S_{i,j}) = \psi(C_j)/\psi(S_{i,j})$. Note that these quantities are given by a direct application of Theorems 2 and 3, for instance we have

$$\mathbb{P}(C_j|C_i) = \frac{\Phi_j(C_j)}{\delta_{j\to i}(S_{i,j})}\prod_{k\neq i,j-k}\delta_{k\to j}(S_{j,k}).$$

□

1.5 Underflow issues and logarithmic computations

Dealing with product of probabilities, one may inevitably encounter numerical underflow issues when the number of potentials grows large. The distributivity property of the sum law precludes one to replace product of potentials by sum of log potentials in Equation (1.8). In this section, we explain a method for overcoming underflow issues which consists in rescaling messages during their implementation in order to keep them in an acceptable range.

Let $J = (C = \{C_1, \dots, C_m\}, \mathcal{E})$ be a JT defined by a variable elimination over a factor graph, we define for each edge $C_i \to C_j$ the rescaled message $\widetilde{\delta}_{i\to j}$ and the logarithmic factor $L_{i\to j}$ to be

$$\widetilde{\delta}_{i\to j}(S_{i,j}) = \frac{\delta_{i\to j}(S_{i,j})}{\sum_{S_{i,j}} \delta_{i\to j}(S_{i,j})} \qquad \text{and} \qquad L_{i\to j} = \log \sum_{S_{i,j}} \delta_{i\to j}(S_{i,j}).$$

A recursive implementation similar to the one seen in Section 1.4.2 adapted to rescaled messages and logarithmic factors is rendered possible with the following theorem:

Theorem 4. *For all $i, j \in \{1, \dots, m\}, j \neq i$ such that $C_i \to C_j$,*

$$\widetilde{\delta}_{i\to j}(S_{i,j}) = \frac{\displaystyle\sum_{C_i \setminus S_{i,j}} \Phi_i(C_i) \prod_{\substack{k\notin\{i,j\}\\ k-i}} \widetilde{\delta}_{k\to i}(S_{i,k})}{\displaystyle\sum_{C_i} \Phi_i(C_i) \prod_{\substack{k\notin\{i,j\}\\ k-i}} \widetilde{\delta}_{k\to i}(S_{i,k})}$$

and

$$L_{i\to j} = \sum_{\substack{k\notin\{i,j\}\\ k-i}} L_{k\to i} + \log \sum_{C_i} \Phi_i(C_i) \prod_{\substack{k\notin\{i,j\}\\ k-i}} \widetilde{\delta}_{k\to i}(S_{i,k})$$

Proof. Multiplying both numerator and denominator by $\left(\exp \displaystyle\sum_{\substack{k\notin\{i,j\}\\ k-i}} L_{k\to i}\right)$ in the first equation proves it and, by induction, for the second equation. □

And finally, the marginal of a separator or a clique can be computed using solely rescaled messages and logarithmic factors. Indeed, adapting Theorems 2 and 3, for each separator $S_{i,j}$ and each clique C_i, we have respectively

$$\psi(S_{i,j}) = \widetilde{\delta}_{i\to j}(S_{i,j})\widetilde{\delta}_{j\to i}(S_{i,j}) \exp\left(L_{i\to j} + L_{j\to i}\right)$$

and respectively

$$\psi(C_i) = \Phi_i(C_i) \prod_{j,i-j} \widetilde{\delta}_{j\to i}(S_{i,j}) \exp\left(\sum_{j,i-j} L_{j\to i}\right).$$

1.6 Computational shortcuts

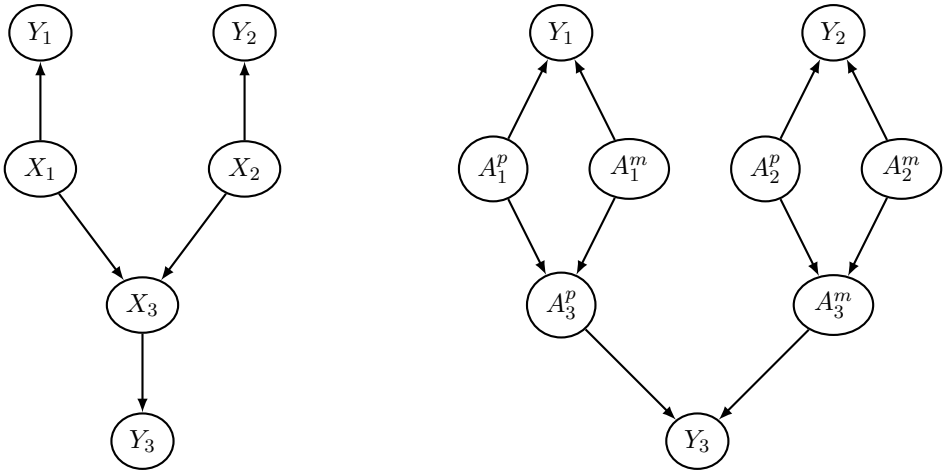
Let $\phi = \{\phi_a\}_{a=1,\dots,p}$ be a set of potentials, $X = \{X_1, \dots, X_n\} = \cup_{a=1}^{p} \text{Scope}(\phi_a)$ and σ be an elimination ordering over X, we have seen in Section 1.4.2 that the algorithmic complexity of an exact inference over X is of order $\mathcal{O}\left(\sum_{i=1}^{m} \prod_{X_u \in C_i} |\mathcal{X}_u|\right)$ where $C = \{C_1, \dots, C_m\}$ is the set of cliques of a JT defined by $\text{VE}(\phi, X, \sigma)$. Therefore, the fewer component edges in the original factor graph H_ϕ and the lower the cardinality of its component variables, the lower the expected complexity. We introduce in this section several computational shortcuts for reducing the problem from the original factor graph (see Shachter, 1988; Fishelson and Geiger, 2002; Lauritzen and Sheehan, 2003). The idea is to build a set of minimal potentials in terms of number of variables and their cardinality composing their scopes.

Splitting potentials. Because the resulting complexity is linear in the number of cliques but exponential in the cardinality of their component variables, one may take advantage of splitting variables into variables of smaller sets of values. As a result, more variables are involved but their sets of values is reduced. Consider for instance a trio composed of a father, a mother and a child respectively associated with indexes 1, 2 and 3 and their genotypes regarding a single biallelic gene and phenotypes coded by that gene. For all $u \in \{1, 2, 3\}$, let X_u denote the genotype of individual u which takes its values in $\{00, 01, 10, 11\}$ and Y_u be his/her phenotype. A DAG involving genotypes and phenotypes is pictured in Figure 1.3a. One may take advantage of splitting the genotypes into their component paternal and maternal alleles, each of them being a binary variable, denoted A_u^p and A_u^m. The resulting DAG is represented in Figure 1.3b. That trick can lead to a high computational boost for a future inference in particular when several members and/or several genes are involved.

Reducing potentials. Entering an evidence in a BN usually renders a subset of its potentials sparse, in particular those which contain variables assigned a value or a subset of values in their scope. For instance entering $\text{ev} = \{X_1 = 0, X_5 = 0, X_7 \in \{1, 2\}\}$ in $\mathcal{B}^{\text{toy}}$ leads to sparse potentials ϕ_{X_1}, ϕ_{X_2}, ϕ_{X_5}, ϕ_{X_6} and ϕ_{X_7} as previously seen in Table 1.1. Reducing a potential consists in suppressing its lines and/or rows of zeros. Note in particular that a variable being assigned a single value is even removed from scopes and in particular any potential whose scope is simply such variable becomes a constant. Table 1.2 gives the set of reduced potentials in $\mathcal{B}^{\text{toy}}$ with evidence $\text{ev} = \{X_1 = 0, X_5 = 0, X_7 \in \{1, 2\}\}$.

We also propose, in Figure 1.4, a graphical illustration of the subsequent removal of nodes and edges of factor graphs induced by sets of potentials reduced respectively with evidence $\{X_7 = 0\}$ and $\{X_1 = 0, X_5 = 0, X_7 \in \{1, 2\}\}$.

Pruning Pruning consists in removing nodes and their descendants if they all are unobserved. Let X_v be a node with no child in BN $\mathcal{B}$, the only CPD it belongs to is $\mathbb{P}(X_v | X_{\text{pa}(v)})$. Let $\text{ev} = \cap_{u \in E} \{X_u \in \mathcal{X}_u^* \subset \mathcal{X}_u\}$ be an evidence for $\mathcal{B}$, if $v \notin E$, entering ev does not modify its values in $\mathbb{P}(X_v | X_{\text{pa}(v)})$ and $\sum_{X_v} \mathbb{P}(X_v | X_{\text{pa}(v)}) = 1$.



(a) DAG involving genotypes and phenotypes of a trio

(b) DAG involving alleles and phenotypes of a trio

Figure 1.3: Two versions of a DAG over which the joint distribution of genotypes and phenotypes of family members factorizes. In the allele version, component variables are of smaller sets of values.

0.6

(a) $\phi_1(\emptyset)$

$X_2 = 0$	$X_2 = 1$	$X_2 = 2$
0.7	0.2	0.1

(b) $\phi_2(X_2)$

$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
0.6	0.3	0.1

(c) $\phi_3(X_3)$

	$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
$X_4 = 0$	0.7	0.6	0.2
$X_4 = 1$	0.2	0.2	0.3
$X_4 = 2$	0.1	0.2	0.5

(d) $\phi_4(X_3, X_4)$

	$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
$X_2 = 0$	0.9	0.7	0.3
$X_2 = 1$	0.7	0.2	0.1
$X_2 = 2$	0.3	0.1	0.05

(e) $\phi_5(X_2, X_3)$

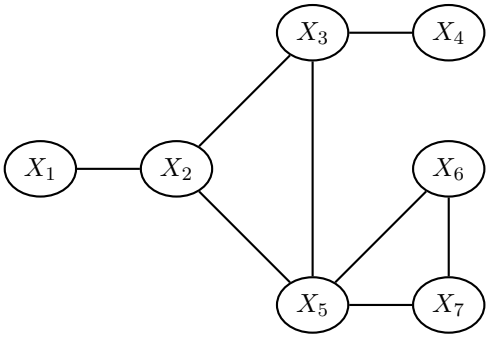
$X_6 = 0$	$X_6 = 1$	$X_6 = 2$
0.7	0.2	0.1

(f) $\phi_6(X_6)$

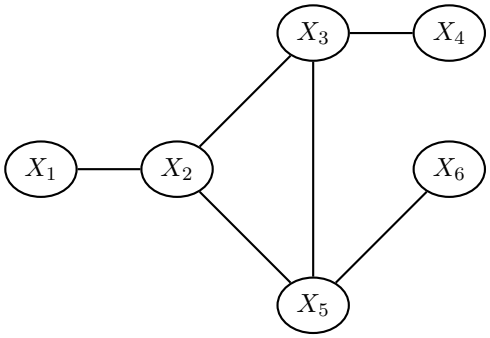
	$X_6 = 0$	$X_6 = 1$	$X_6 = 2$
$X_7 = 1$	0.05	0.2	0.5
$X_7 = 2$	0.05	0.1	0.2

(g) $\phi_7(X_6, X_7)$

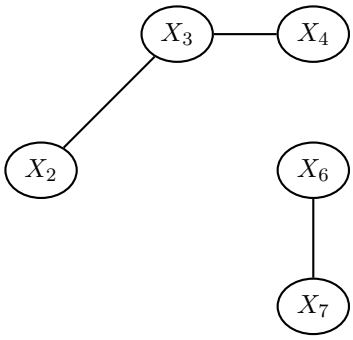
Table 1.2: Potentials associated with $\mathcal{B}^{\text{toy}}$ after entering ev $= \{X_1 = 0, X_5 = 0, X_7 \in \{1, 2\}\}$ and reducing.



(a) Moral graph of G^{toy} ($H_{\phi\{\emptyset\}}$)



(b) Factor graph $H_{\phi\{X_7=0\}}$



(c) Factor graph $H_{\phi\{X_1=0,X_5=0,X_7\in\{1,2\}\}}$

Figure 1.4: Factor graphs induced by different sets of potentials ϕ^{ev} defined as the set of reduced CPDs in $\mathcal{B}_{\text{toy}}$ after entering ev.

	$X_5 = 0, X_6 = 0$	$X_5 = 1, X_6 = 0$	$X_5 = 2, X_6 = 0$
$X_7 = 0$	1.0	0.0	0.0
$X_7 = 1$	0.0	0.5	0.0
$X_7 = 2$	0.0	0.5	1.0
	$X_5 = 0, X_6 = 1$	$X_5 = 1, X_6 = 1$	$X_5 = 1, X_6 = 2$
$X_7 = 0$	0.0	0.0	0.0
$X_7 = 1$	1.0	0.0	0.5
$X_7 = 2$	0.0	1.0	0.5
	$X_5 = 0, X_6 = 2$	$X_5 = 1, X_6 = 2$	$X_5 = 2, X_6 = 2$
$X_7 = 0$	0.0	0.0	0.0
$X_7 = 1$	0.0	0.5	0.0
$X_7 = 2$	1.0	0.5	1.0

Table 1.3: CPD $\mathbb{P}(X_7|X_5, X_6)$ in $\mathcal{B}^{\text{det.}}$

The CPD $\mathbb{P}(X_v|X_{\text{pa}(v)})$ can simply be removed from the set of CPDs, therefore, X_v removed from the DAG along with edges towards it. Once a terminal unobserved node is removed, reiterating the operation subsequently leads to removing all sets of nodes along with their descendants if they all are unobserved along with their associated CPDs. A set of nodes $X_V \subset X$ is said to *barren* relative to $X \setminus X_V$ if X_V is irrelevant to an inference regarding nodes in $X \setminus X_V$. Of course, removing a variable implies that no query can later be made for it and one may prefer not to.

Forcing. The introduction of a hard evidence in a BN containing deterministic CPDs may lead to extra unnecessary dependencies. Let us consider a modified version of $\mathcal{B}^{\text{toy}}$ denoted $\mathcal{B}^{\text{det.}}$ over the same DAG and with same CPDs except $\mathbb{P}(X_7|X_5, X_6)$ which modified version is given in Table 1.3. Consider the hard evidence ev $= \{X_7 = 0\}$, then X_5 and X_6 deterministically take value 0. They can be included in the evidence as if they where observed leading to the evidence $\{X_5 = 0, X_6 = 0, X_7 = 0\}$. Entering that new evidence reduces the state space of potentials associated with $\mathbb{P}(X_7|X_5, X_6)$ but also $\mathbb{P}(X_5|X_2, X_3)$ and $\mathbb{P}(X_6|X_5)$. The set of resulting potentials and the factor graph induced by that set are represented in Figure 1.5. Note furthermore that the resulting potentials ϕ_7 is neutral and can simply be removed.

An efficient method to detect forcing options is proposed by Cottingham Jr et al. (1993). Because potentials in that context contain probabilities and we cannot get zeros by adding nonzero values, the authors suggest to perform a boolean propagation of the evidence where real-valued potentials are replaced by boolean potentials such that each value 0.0 (respectively nonzero values) is replaced by the boolean value FALSE (respectively TRUE). Perform the sum-product algorithm with boolean potentials where the sum operator (respectively the product operator) is replaced by OR (respectively AND) operator, marginalize each variable and replace a TRUE result by the corresponding original real value. Because a boolean propagation is a lot faster than a propagation over real values and conventional sum and product,

$X_1 = 0$	$X_1 = 1$	$X_1 = 2$
0.6	0.3	0.1

(a) $\phi_1(X_1)$

	$X_1 = 0$	$X_1 = 1$	$X_1 = 2$
$X_2 = 0$	0.7	0.6	0.2
$X_2 = 1$	0.2	0.2	0.3
$X_2 = 2$	0.1	0.2	0.5

(b) $\phi_2(X_1, X_2)$

$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
0.6	0.3	0.1

(c) $\phi_3(X_3)$

	$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
$X_4 = 0$	0.7	0.6	0.2
$X_4 = 1$	0.2	0.2	0.3
$X_4 = 2$	0.1	0.2	0.5

(d) $\phi_4(X_3, X_4)$

	$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
$X_2 = 0$	0.9	0.7	0.3
$X_2 = 1$	0.7	0.2	0.1
$X_2 = 2$	0.3	0.1	0.05

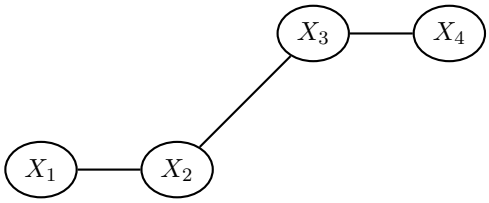
(e) $\phi_5(X_2, X_3)$

0.7

(f) $\phi_6(\emptyset)$

1.0

(g) $\phi_7(\emptyset)$



(h) Factor graph

Figure 1.5: Set $\phi = \{\phi_1, \ldots, \phi_7\}$ of potentials associated with $\mathcal{B}^{\text{det}\cdot}$ after entering ev $= \{X_7 = 0\}$ and forcing (tables on the top) and factor graph induced by ϕ (at the bottom).

performing such propagation before the sum-product algorithm can lead to dramatic gain in computational complexity.

Remark. Computational shortcuts may lead to disconnected factor graphs and/or empty scopes of potentials. In the previous sections we detailed the sum-product algorithm for connected factor graphs and non empty scopes but the extension to disconnected factor graphs and empty scopes is straightforward. Indeed, let $\phi = \{\phi_1, \dots, \phi_p\} = \{\phi_A, \phi_B, \phi_D\}$ be a set of potentials such that $A \sqcup B \sqcup D = \{1, \dots, p\}$ and $\phi_A = \{\phi_a\}_{a \in A}$, $\phi_B = \{\phi_b\}_{b \in B}$, $\phi_D = \{\phi_d\}_{d \in D}$. Let $X_{S_A} = \cup_{a \in A} \text{Scope}(\phi_a)$, $X_{S_B} = \cup_{b \in B} \text{Scope}(\phi_b)$ and $X_{S_D} = \cup_{d \in D} \text{Scope}(\phi_d)$, we assume that $X_{S_D} = \emptyset$ and $X = X_{S_A} \sqcup X_{S_B}$ and therefore we have

$$\sum_X \prod_{c \in \{A,B,D\}} \phi_c(X_{S_c}) = \left(\prod_{d \in D} \phi_d(\emptyset) \right) \left(\sum_{X_{S_A}} \prod_{a \in A} \phi_a(X_{S_a}) \right) \left(\sum_{X_{S_B}} \prod_{b \in B} \phi_b(X_{S_b}) \right)$$

where the scope of a potential ϕ_c for $c \in \{1, \dots, p\}$ is denoted X_{S_c}. For a given query, each factor graph H_{ϕ_A} and H_{ϕ_B} can be treated separately with the sum-product algorithm and results multiplied together and multiplied by the product of constants $\prod_{d \in D} \phi_d(\emptyset)$.

1.7 Particular cases

We introduce in that section some particular cases that will be encountered throughout the thesis.

1.7.1 Bayesian network

The particular case of a Bayesian network has previously been mentioned in Sections 1.1.1 and 1.1.2. We have $\psi(X) = \mathbb{P}(X, \text{ev})$ where $X = \cup_{a=1}^{p} \text{Scope}(\phi_a)$ such that $\phi = \{\phi_a\}_{a=1,\dots,p}$ is the set of potentials obtained from CPDs after entering an evidence ev and/or applying computational shortcuts.

Let us propose an illustration over a particular example. We return to the BN $\mathcal{B}^{\text{toy}}$ and consider the evidence $\text{ev} = \{X_7 = 0\}$. Let $\phi^{\{X_7=0\}} = \{\phi_1, \dots, \phi_7\}$ be the set of potentials obtained after entering ev and reducing, the set $\phi^{\{X_7=0\}}$ is given in Table 1.4 as well as a graphical representation of the factor graph $H_{\phi^{\{X_7=0\}}}$ in Figure 1.4b. Note that $H_{\phi^{\{X_7=0\}}}$ is chordal and $\sigma = (X_6, X_5, X_4, X_3, X_2, X_1)$ is a perfect elimination ordering over $X = \{X_1, \dots, X_6\}$ for $H_{\phi^{\{X_7=0\}}}$ as it corresponds to the subsequent removal of simplicial nodes. The variable elimination $\text{VE}(\phi^{\{X_7=0\}}, X, \sigma)$ defines the JT represented in Figure 1.6. For all $i \in \{1, \dots, 5\}$, let $\Phi_i(C_i)$ be the clique potential as defined in Definition 1.9 and let $\{\delta_{i \to j}\}_{i,j \in \{1,\dots,5\}, C_i \to C_j}$ be the set of messages implemented as detailed in Section 1.4 with a time complexity below $2 \times 5 \times 3^3$, any marginal posterior probability can be computed picking either a separator, a clique or a path containing the variables in the query (with a complexity exponential in the number of variables picked) and applying Theorem 2 or 3

$X_1 = 0$	$X_1 = 1$	$X_1 = 2$
0.6	0.3	0.1

(a) $\phi_1(X_1)$

	$X_1 = 0$	$X_1 = 1$	$X_1 = 2$
$X_2 = 0$	0.7	0.6	0.2
$X_2 = 1$	0.2	0.2	0.3
$X_2 = 2$	0.1	0.2	0.5

(b) $\phi_2(X_1, X_2)$

$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
0.6	0.3	0.1

(c) $\phi_3(X_3)$

	$X_3 = 0$	$X_3 = 1$	$X_3 = 2$
$X_4 = 0$	0.7	0.6	0.2
$X_4 = 1$	0.2	0.2	0.3
$X_4 = 2$	0.1	0.2	0.5

(d) $\phi_4(X_3, X_4)$

	$X_2 = 0, X_3 = 0$	$X_2 = 1, X_3 = 0$	$X_2 = 2, X_3 = 0$
$X_5 = 0$	0.9	0.7	0.3
$X_5 = 1$	0.05	0.2	0.5
$X_5 = 2$	0.05	0.1	0.2

	$X_2 = 0, X_3 = 1$	$X_2 = 1, X_3 = 1$	$X_2 = 1, X_3 = 2$
$X_5 = 0$	0.7	0.2	0.1
$X_5 = 1$	0.2	0.5	0.4
$X_5 = 2$	0.1	0.3	0.5

	$X_2 = 0, X_3 = 2$	$X_2 = 1, X_3 = 2$	$X_2 = 2, X_3 = 2$
$X_5 = 0$	0.3	0.1	0.05
$X_5 = 1$	0.5	0.4	0.05
$X_5 = 2$	0.2	0.5	0.9

(e) $\phi_5(X_2, X_3, X_5)$

	$X_5 = 0$	$X_5 = 1$	$X_5 = 2$
$X_6 = 0$	0.7	0.6	0.2
$X_6 = 1$	0.2	0.2	0.3
$X_6 = 2$	0.1	0.2	0.5

(f) $\phi_6(X_5, X_6)$

	$X_6 = 0$	$X_6 = 1$	$X_6 = 2$
$X_5 = 0$	0.9	0.7	0.3
$X_5 = 1$	0.7	0.2	0.1
$X_5 = 2$	0.3	0.1	0.05

(g) $\phi_7(X_5, X_6)$

Table 1.4: Potentials in $\mathcal{B}^{\text{toy}}$ after entering ev $= \{X_7 = 0\}$ and reducing.

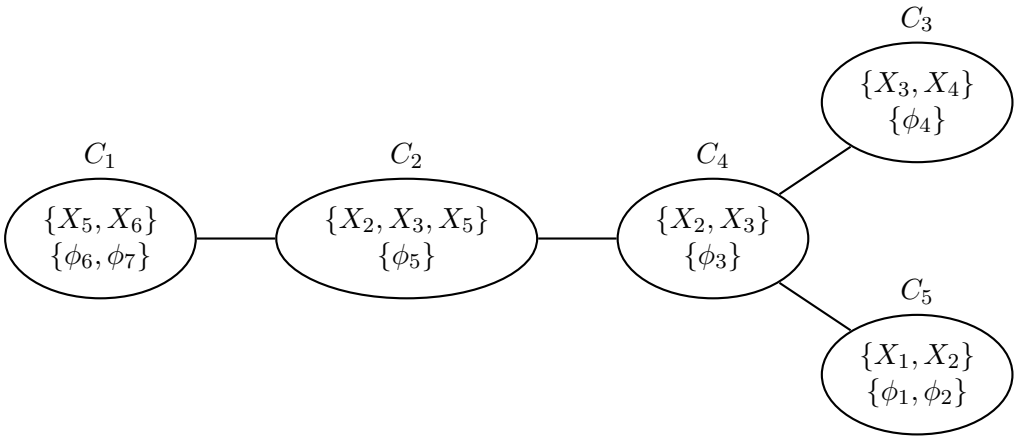


Figure 1.6: Junction-tree defined by $\mathrm{VE}(\phi^{\{X_7=0\}}, X = \{X_1, \dots, X_6\}, \sigma = (X_6, \dots, X_1))$ in $\mathcal{B}^{\mathrm{toy}}$ with injected potentials per clique.

appropriately. For instance we have

$$\mathbb{P}(X_1|\mathrm{ev}) = \frac{1}{\mathcal{Z}} \sum_{X_2} \Phi_5(X_1, X_2)\delta_{4\to 5}(X_2) \quad \text{and} \quad \mathbb{P}(X_3|\mathrm{ev}) = \frac{1}{\mathcal{Z}} \delta_{3\to 4}(S_{3,4})\delta_{4\to 3}(S_{3,4})$$

with $\mathcal{Z} = \mathbb{P}(\mathrm{ev}) = \mathbb{P}(X_7 = 0) = \sum_{X_3} \delta_{3\to 4}(S_{3,4})\delta_{4\to 3}(S_{3,4})$.

1.7.2 Markov chain

Another particular example is given by a Markov chain $X = (X_1, \dots, X_n)$ such that, for all $i \in \{1, \dots, n\}$, $\mathcal{X}_i$ is the set of values taken by X_i and

$$\mathbb{P}(X_1, \dots, X_n) = \mathbb{P}(X_1) \prod_{i=2}^{n} \mathbb{P}(X_i|X_{i-1}).$$

We define an evidence $\mathrm{ev} = \{\cap_{i \in E \subset \{1,\dots,n\}} X_i \in \mathcal{X}_i^* \subset \mathcal{X}_i\}$ and for all $i \in \{1, \dots, n\} \setminus E$, let $\mathcal{X}_i^* = \mathcal{X}_i$. We introduce

$$\psi(X) = \mathbb{P}(X, \mathrm{ev}) = \phi_1(X_1) \prod_{i=2}^{n} \phi_i(X_{i-1}, X_i)$$

where

$$\phi_1(X_1) = \mathbb{1}_{X_1 \in \mathcal{X}_1^*} \mathbb{P}(X_1)$$

and, for all $i \in \{2, \dots, n\}$,

$$\phi_i(X_{i-1}, X_i) = \mathbb{1}_{X_i \in \mathcal{X}_i^*} \mathbb{P}(X_i|X_{i-1}).$$

Let $\phi = \{\phi_i\}_{i=1,\dots,n}$, the elimination ordering $\sigma = (X_1, \dots, X_n)$ is perfect for H_ϕ and the junction-tree $J = (C = (C_1, \dots, C_{n+1}), \mathcal{F})$ defined by $\mathrm{VE}(\phi, X, \sigma)$ is given by

- $C_1 = X_1$, for all $i \in \{2, \dots, n\}$, $C_i = \{X_{i-1}, X_i\}$ and $C_{n+1} = X_n$

- For all $i \in \{1, \dots, n\}$, $C_i^* = \phi_i$ and $C_{n+1}^* = \emptyset$
- $\mathcal{F} = \{(C_i, C_{i+1})\}_{i=1,\dots,n}$.

We define quantities respectively called *foward* and *backward messages* as

$$F_i(X_i) \stackrel{\text{def}}{=} \mathbb{P}(X_1 \in \mathcal{X}_1^*, \dots, X_{i-1} \in \mathcal{X}_{i-1}^*, X_i)$$

and

$$B_i(X_i) \stackrel{\text{def}}{=} \mathbb{P}(X_{i+1} \in \mathcal{X}_{i+1}^*, \dots, X_n \in \mathcal{X}_n^* | X_i)$$

which can be recursively computed with a forward and a backward pass choosing $C_{n+1} = X_n$ as the root. The forward pass is initialized by, for all $x \in \mathcal{X}_1^*$, $F_1(x) = \phi_1(x)$ and for all $i = 2, \dots, n$, for all $y \in \mathcal{X}_i^*$,

$$F_i(y) = \sum_{x \in \mathcal{X}_{i-1}^*} F_{i-1}(x)\phi_i(x, y).$$

The backward pass is initialized by, for all $x \in \mathcal{X}_n^*$, $B_n(x) = 1$ and for all $i = n, \dots, 2$, for all $x \in \mathcal{X}_{i-1}^*$,

$$B_{i-1}(x) = \sum_{y \in \mathcal{X}_i^*} \phi_i(x, y)B_i(y).$$

Note that for all $i \in \{1, \dots, n-1\}$, F_i (respectively B_i) is precisely the message $\delta_{i \to i+1}$ (respectively $\delta_{i+1 \to i}$ defined in Definition 9). In this framework, σ is perfect for the (chordal) factor graph which is a sequence, so is the JT J defined by $\text{VE}(\phi, X, \sigma)$. Because each clique $C_i \in \{C_1, \dots, C_n\}$ sends and receives a unique message from $C_{\text{to}(i)}$ respectively during the inward and the outward pass, the sum-product algorithm is usually called *forward-backward* algorithm in the framework of a Markov chain.

One can finally infer marginal posterior probabilities of individual states as well as transition probabilities with a direct application of Theorem 2 and 3. Indeed, for all $i \in \{1, \dots, n\}$, we have

$$\mathbb{P}(X_i|\text{ev}) = \frac{1}{\mathcal{Z}}\psi(X_i) = \frac{1}{\mathcal{Z}}F_i(X_i)B_i(X_i) \quad \text{where} \quad \mathcal{Z} = \sum_{X_i} F_i(X_i)B_i(X_i)$$

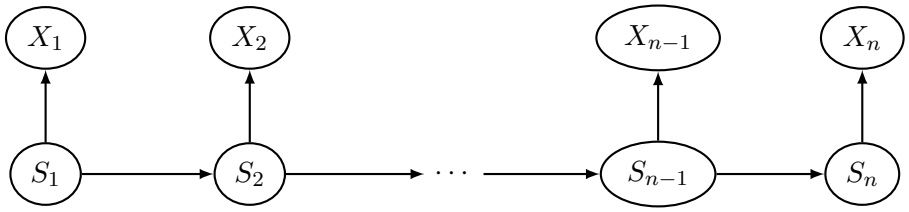
and

$$\mathbb{P}(X_i|X_{i-1}, \text{ev}) = \frac{\psi(X_{i-1}, X_i)}{\psi(X_{i-1})} = \frac{\phi_i(X_{i-1}, X_i)B_i(X_i)}{B_{i-1}(X_{i-1})}.$$

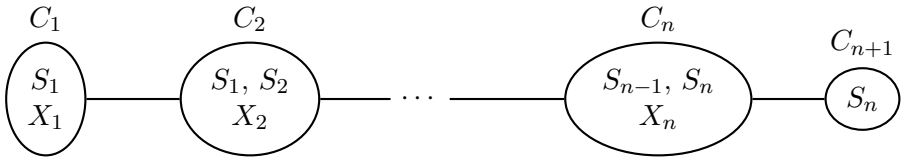
1.7.3 Hidden Markov model

In the particular case of a Hidden Markov Model (HMM), the recursion shares many similarities with Markov chains. We consider an HMM (X, S) represented in Figure 1.7a such that $S = (S_1, \dots, S_n) \in \mathcal{S}^n$ is the set of hidden variables and $X = (X_1, \dots, X_n) \in \mathcal{X}^n$ with

$$\mathbb{P}(X, S) = \mu(S_1)\eta(S_1, X_1) \prod_{i=2}^{n} \pi(S_{i-1}, S_i)\eta(S_i, X_i)$$



(a) Example of a hidden Markov model



(b) Junction-tree

Figure 1.7: Example of a hidden Markov model over (X, S) and junction-tree defined by $\mathrm{VE}(\phi, \{X, S\}, \sigma)$ with $\sigma = (X_1, S_1, \ldots, X_n, S_n)$.

where μ, π and η are respectively called the initial distribution, the transition and emission probabilities. Let $\mathrm{ev} = \cap_{i \in E}\{X_i \in \mathcal{X}_i^* \subset \mathcal{X}\}$, $E \subset \{1, \ldots, n\}$, be an evidence for the HMM and for all $i \notin E$, let $\mathcal{X}_i^* = \mathcal{X}_i$, we define

$$\psi(X, S) = \mathbb{P}(X, S, \mathrm{ev}) = \phi_1(S_1, X_1) \prod_{i=2}^{n} \phi_i(S_{i-1}, S_i, X_i)$$

such that, for all $r, s \in \mathcal{S}$ and $x \in \mathcal{X}$,

$$\phi_1(s, x) = \mathbb{1}_{x \in \mathcal{X}_1^*} \mu(s) \eta(s, x) \quad \text{and} \quad \phi_i(r, s, x) = \mathbb{1}_{x \in \mathcal{X}_i^*} \pi(r, s) \eta(s, x).$$

We introduce $\phi = \{\phi_i\}_{i=1,\ldots,n}$ and the elimination ordering $\sigma = (X_1, S_1, \ldots, X_n, S_n)$ which is perfect for H_ϕ. $\mathrm{VE}(\phi, \{X, S\}, \sigma)$ defines the JT $J = (C = (C_1, \ldots, C_{n+1}), \mathcal{F})$ represented in Figure 1.7b where

- $C_1 = \{S_1, X_1\}$, for all $i \in \{2, \ldots, n\}$, $C_i = \{S_{i-1}, S_i, X_i\}$ and $C_{n+1} = S_n$,
- For all $i \in \{1, \ldots, n\}$, $C_i^* = \phi_i$ and $C_{n+1}^* = \emptyset$,
- $\mathcal{F} = \{(C_i, C_{i+1})\}_{i=1,\ldots,n}$.

We define quantities respectively called *forward* and *backward messages* as

$$F_i(S_i) \overset{\text{def}}{=} \mathbb{P}(X_1 \in \mathcal{X}_1^*, \ldots, X_i \in \mathcal{X}_i^*, S_i)$$

and

$$B_i(S_i) \overset{\text{def}}{=} \mathbb{P}(X_{i+1} \in \mathcal{X}_{i+1}^*, \ldots, X_n \in \mathcal{X}_n^* | S_i)$$

which can be recursively computed with a forward and backward pass choosing $C_{n+1} = \{S_n\}$ as the root. For all $s \in \mathcal{S}$ and $x \in \mathcal{X}$, the forward pass is initialized with $F_1(s) = \sum_{x \in \mathcal{X}} \phi_1(s, x)$ and for $i = 2, \ldots, n$,

$$F_i(s) = \sum_{x \in \mathcal{X}} \sum_{r \in \mathcal{S}} F_{i-1}(r) \phi_i(r, s, x)$$

For all $r \in \mathcal{S}$, the backward pass is initialized with $B_n(r) = 1$ and, for all $i = n, \dots, 2$, $x \in \mathcal{X}$,

$$B_{i-1}(r) = \sum_{x \in \mathcal{X}} \sum_{s \in \mathcal{S}} \phi_i(r, s, x) B_i(s).$$

Note again that for all $i \in \{1, \dots, n-1\}$, F_i (respectively B_i) is precisely the message $\delta_{i \to i+1}$ (respectively $\delta_{i+1 \to i}$ defined in Definition 9) and J is a sequence. Consequently, for same reasons as those mentioned for a Markov chain, the message passing algorithm is also usually called the *forward-backward* algorithm in the framework of an HMM. At the end of the recursion, applying Theorem 2 and/or 3 appropriately, one can infer $\mathbb{P}(\text{ev})$ or the marginal of a chosen hidden variable. Indeed we have, for all $i \in \{1, \dots, n\}$, $\psi(S_i) = \mathbb{P}(S_i, \text{ev}) = F_i(S_i) B_i(S_i)$ and $\psi(S_{i-1}, S_i, X_i) = \mathbb{P}(S_{i-1}, S_i, X_i, \text{ev}) = F_{i-1}(S_{i-1}) \phi_i(S_{i-1}, S_i, X_i) B_i(S_i)$. Therefore, one can therefore choose any index $i \in \{1, \dots, n+1\}$ to compute

$$\mathbb{P}(\text{ev}) = \mathcal{Z} = \sum_{s \in \mathcal{S}} F_i(s) B_i(s).$$

In particular, if no other inference is needed, choosing $i = n+1$ allows one for avoiding the computation of backward messages. Moreover, one can infer, at a chosen index i, individual posterior state probabilities:

$$\mathbb{P}(S_i | \text{ev}) = \frac{1}{\mathcal{Z}} F_i(S_i) B_i(S_i)$$

as well as transition probabilities:

$$\mathbb{P}(S_i | S_{i-1}, \text{ev}) = \frac{\psi(S_{i-1}, S_i)}{\psi(S_{i-1})} = \frac{\sum_{x \in \mathcal{X}} \phi_i(S_{i-1}, S_i, X_i = x) B_i(S_i)}{B_{i-1}(S_{i-1})}$$

and emission probabilities (for which one can avoid the backward pass of the algorithm if no other inference is needed):

$$\mathbb{P}(X_i | S_i, \text{ev}) = \frac{\psi(X_i, S_i)}{\psi(S_i)} = \frac{\sum_{r \in \mathcal{S}} F_{i-1}(r) \phi_i(r, S_i, X_i)}{F_i(S_i)}$$

1.8 MAP and marginal MAP inference

In this section we detail an algorithm similar to the sum-product called *max-product* (or *max-sum*) for performing inference of the Maximum A Posteriori (MAP) also called Most Probable Explanation (MPE).

1.8.1 MAP inference

A MAP inference aims at determining the most likely values of latent variables in a BN $\mathcal{B} = (G = (\{X_1, \dots, X_n\}, \mathcal{E}), \mathbb{P})$ given a hard evidence $\text{ev} = \{X_E = x_e\}$ where $X_E \subset X$ and x_e is a vector of observed values taken by X_E. The MAP conditional on ev is defined as

$$\text{MAP}(X_U | \text{ev}) \stackrel{\text{def}}{=} \arg\max_{X_U} \mathbb{P}(X_U | \text{ev}) = \arg\max_{X_U} \mathbb{P}(X_U, \text{ev}) \tag{1.10}$$

where $X_U = X \setminus X_E$ and may have more than one solution.

Usually we have $\text{MAP}(X_U|\text{ev}) \neq \cap_{u \in U} \text{MAP}(X_u|\text{ev})$ as illustrated bellow over a simple BN denoted $\mathcal{B}^s$ composed of three binary variables A, B, C over the DAG $A \to B \to C$ and CPDs listed in Table 1.5.

$A=0$	$A=1$
0.5	0.5

(a) $\mathbb{P}(A)$

	$A=0$	$A=1$
$B=0$	0.3	0.2
$B=1$	0.7	0.8

(b) $\mathbb{P}(B|A)$

	$B=0$	$B=1$
$C=0$	0.8	0.4
$C=1$	0.2	0.6

(c) $\mathbb{P}(C|B)$

Table 1.5: LPDs in the BN $\mathcal{B}^s$ whose DAG is $A \to B \to C$.

Let us enter the evidence $\{A = 0\}$, the marginal probability $\mathbb{P}(B|A = 0)$ as well as $\mathbb{P}(C|A = 0) = \sum_B \mathbb{P}(B|A = 0)\mathbb{P}(C|B)$ and the joint probability $\mathbb{P}(B, C|A = 0)$ are reported in Table 1.6. Note that $\text{MAP}(B|A = 0) = 1$, $\text{MAP}(C|A = 0) = 0$ and $\text{MAP}(B, C|A = 0) = (1, 1) \neq (\text{MAP}(B|A = 0), \text{MAP}(C|A = 0))$.

$B=0$	$B=1$
0.3	0.7

(a) $\mathbb{P}(B|\{A = 0\})$

$C=0$	$C=1$
0.52	0.48

(b) $\mathbb{P}(C|\{A = 0\})$

	$C=0$	$C=1$
$B=0$	0.24	0.06
$B=1$	0.28	0.42

(c) $\mathbb{P}(B, C|\{A = 0\})$

Table 1.6: Posterior marginal probabilities $\mathbb{P}(B|\{A = 0\})$ and $\mathbb{P}(C|\{A = 0\})$ and posterior joint probability $\mathbb{P}(B, C|\{A = 0\})$ in $\mathcal{B}^s$.

Let us define the max marginalization of Y out of a potential ϕ whose scope is $\text{Scope}(\phi) = \{Y, Z\}$ to be $\max_Y \phi(Y, Z) = \widetilde{\phi}(Z)$ such that $\widetilde{\phi}(Z) = \max \phi(Y = y, Z = z)$ for each value y taken by Y. In other words $\max_Y \phi(Y, Z)$ is a potential which contains the most likely assignment or joint assignments z for each value y.

We directly deduce from Equation (1.10) that

$$\text{MAP}(X_U|\text{ev}) = \arg\max_{X_U} \psi(X_U) \quad \text{where} \quad \psi(X_U) = \prod_{a=1}^{p} \phi_a(X_{S_a})$$

where $\phi = \{\phi_a\}_{a=1,\dots,p}$ is the set of potentials obtained from CPDs in the BN after entering ev and/or applying computational shortcuts. Therefore algorithms similar to the sum-product variable elimination and the sum-product algorithm developed in previous sections where the $\sum$ is replaced by the max operator can be used for reducing the complexity to compute $\text{MAP}(X_U|\text{ev})$. Indeed a variable elimination as detailed in Algorithm 1 where the $\sum$ is replaced by max operator allows one for computing $\max_{X_U} \prod_{a=1}^{p} \phi_a(X_{S_a})$ in same order complexity and an additional traceback procedure starting from the last variable eliminated and finishing with the first returns $\arg\max_{X_U} \prod_{a=1}^{p} \phi_a(X_{S_a})$. The resulting algorithm is called *max-product variable elimination*. Alternatively, one can add a outward pass to the variable elimination, i.e. perform the sum-product algorithm as developed in Algorithm 4

where $\sum$ is replaced by max operator and compute max marginals of chosen sets of variables applying Theorems 2 and 3 where the $\sum$ is replaced by max operator. The resulting algorithm is called *max-product* algorithm. Each pass requires same complexity as those detailed in Section 1.4.2. The operation that consists in building a joint assignment from optimizing local max-marginals is called *decoding* and is unambiguous only when max-marginals are unambiguous such that the solution is a single assignment per variable. We will solely consider unambiguous MAP in the framework of this thesis.

Remark. Underflow issues are easily overcome by replacing the max-product algorithm with its max-sum version where product of potentials are replaced by sum of logarithmic potentials.

1.8.2 Marginal MAP inference

Following the same reasoning, the marginal MAP of a subset of latent variables in $\mathcal{B}$ conditional on $\text{ev} = \{X_E = x_e\}$ is given by

$$\text{MAP}(X_V|\text{ev}) = \arg\max_{X_V} \sum_{X_W} \mathbb{P}(X_V, X_W|\text{ev}) = \arg\max_{X_V} \sum_{X_W} \psi(X_U) \tag{1.11}$$

where $X_V \sqcup X_W = X_U = X \setminus X_E$ and $\psi(X_U) = \prod_{a=1}^{p} \phi_a(X_{S_a})$. As maximization and summation are not commutative, one must first marginalize X_W out of $\psi(X_U)$ before maximizing X_V in the resulting product of potentials which usually leads to higher complexities than inference of posterior probabilities or the MAP. Let us illustrate this point over our example $\mathcal{B}^{\text{toy}}$ with the evidence $\text{ev} = \{X_7 = 0\}$ and inference $\text{MAP}(X_4, X_6|X_7 = 0) = \arg\max_{X_4,X_6} \sum_{X_1,X_2,X_3,X_5} \prod_{u=1}^{n} \phi_u(X_{S_u})$ where $\phi = \{\phi_u\}_{u=1,\dots,7}$ is the set of potentials obtained after entering $\text{ev} = \{X_7 = 0\}$ and reducing its CPDs. The factor graph H_ϕ is represented in Figure 1.4b. An elimination ordering over X to compute $\text{MAP}(X_4, X_6|X_7 = 0)$ with the sum-product followed by the max-product algorithm must satisfy the constraint such that it starts with $X_W = \{X_1, X_2, X_3, X_5\}$ in any ordering and finishes with $X_V = \{X_4, X_6\}$ in any ordering. No perfect elimination ordering over X_W exists and fill-in edges must be added to H_ϕ for instance $X_4 - X_5$ and $X_4 - X_6$.

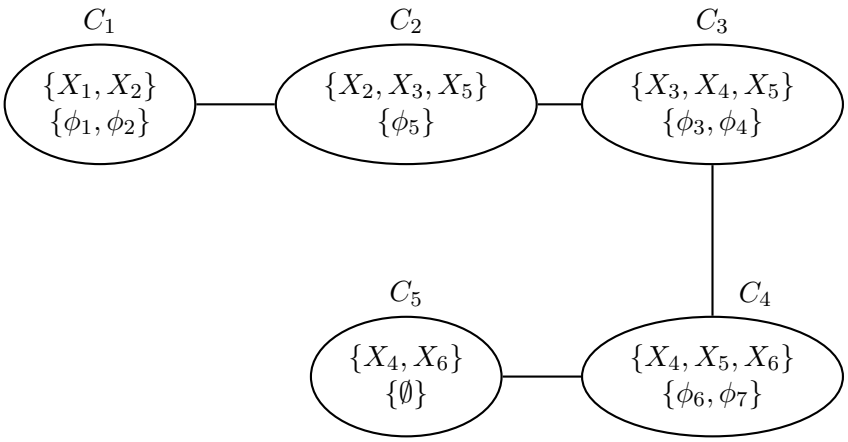


Figure 1.8: Junction-tree defined by $\text{VE}(\phi, X_U, \sigma)$ with injected potentials where ϕ is the set of potentials obtained after entering $\text{ev} = \{X_7 = 0\}$ in CPDs in $\mathcal{B}^{\text{toy}}$ and reducing, $X_U = X \setminus X_7$ and $\sigma = (X_1, X_2, X_3, X_5, X_4, X_6)$ chosen to compute $\text{MAP}(X_4, X_6|\text{ev})$.

Chapter 2

Survival analysis

Sommaire

2.1 General notions in survival analysis

Survival analysis is the study of time-to-event data. A time-to-event data (also called a survival data) is the elapsed time from a time origin until the onset of an event of interest. The random variable of interest is therefore a duration. The event of interest can be of various type (not necessarily death) such as the diagnosis of a disease, a relapse, a recovery, a machine breakdown, a birth, etc. We define the time origin to be the start date of the duration, usually the date the individual starts to be at risk, for instance his/her birth, the start date of an exposure, a disease onset, a surgical operation, etc.

Censoring and truncation. The particularity of time-to-event data is the nature of incomplete data due to censoring and/or truncation. We distinguish three types of censoring:

- Right censoring occurs when an individual do not present the event of interest before his/her date of last news. This is typically the case in a follow-up or cohort study where some individuals quit the study or die during the study or do not present the event before the stop date of the study.

- Left censoring occurs when an individual present the event before it is observed.

- Interval censoring occurs when an individual present the event in between two observation dates. For instance, when a dental cavity forming between two medical appointments.

Truncation is due to sampling bias and occurs when observations are made conditional on a set of events. For instance in a delayed cohort study, individuals who met criteria to enter the study are not enrolled, for instance because of death, are left-truncated.

Right censoring and left truncation are the most commonly encountered causes of missing data. For the rest of the thesis we work in the framework of right censoring and no other censoring nor truncation.

Variables and main functions. For each patient, we introduce the observed duration time $\tilde{T}$, the true duration time T, the censoring time C and the status Δ such that

$$\begin{cases} \tilde{T} = \min(T, C) \\ \Delta = \mathbb{1}_{\{T \leq C\}} \end{cases}$$

where $\mathbb{1}_{\{.\}}$ is the indicatrix function. In most statistical analysis we must assume and verify that T is independent of C. For instance a classical example of non independency between T and C in a cohort study which studies the effect of a treatment is given by individuals who quit the study because of side effects of the treatment. Let F (respectively G) be the cumulative distribution function of T (respectively C), we have

$$\mathcal{P}(\tilde{T} \leq t, \Delta = 1) = \mathbb{E}\big[\mathbb{1}_{\{\tilde{T} \leq t, T \leq C\}}\big] = \int_0^t (1 - G(u))\, dF(u) \leq \mathcal{P}(T \leq t),$$

therefore censored data should not be ignored when studying T.

The distribution of T can be defined by one the the following five functions: for all $t \geq 0$,

- The cumulative distribution function: $F(t) = \mathcal{P}(T < t)$
- The survival function: $S(t) = \mathcal{P}(T \geq t)$
- The density:
$$f(t) = \lim_{\delta t \to 0} \frac{\mathcal{P}(t \leq T < t + \delta t)}{\delta t}$$
- The hazard function:
$$\lambda(t) = \lim_{\delta t \to 0} \frac{\mathcal{P}(t \leq T < t + \delta t | T \geq t)}{\delta t}$$
- The cumulative hazard:
$$\Lambda(t) = \int_0^t \lambda(u) du.$$

Note that the hazard function is a conditional density given by $\lambda(t) = f(t)/S(t)$. All of the above functions are linked together and in particular, denoting the derivative with respect to t with a prime symbol, we have $S'(t) = (1 - F(t))' = -f(t)$, hence, $\lambda(t) = -\big[\log(S(t))\big]'$ and therefore

$$S(t) = \exp\left(-\int_0^t \lambda(u)du\right). \tag{2.1}$$

Non-parametric, semi-parametric and parametric survival analysis. There are three different choices for modeling one of the above functions:

- Non-parametric survival analysis. The main two non-parametric estimators are the Kaplan-Meier estimator of the survival function (Kaplan and Meier, 1958) and the Nelson-Aalen estimator of the cumulative hazard (Nelson, 1969, 1972; Aalen, 1978).

  The Kaplan-Meier estimator of the survival function (Kaplan and Meier, 1958) is a step function given by

  $$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{R_i}\right)$$

  where d_i is the number of individuals presenting the event at time t_i and R_i is the number of individuals at risk of presenting the event at time t_i. The Kaplan-Meier estimator is biased but asymptotically unbiased. The Nelson-Aalen estimator of cumulative hazard (Nelson, 1972; Aalen, 1978) is a step function given by

  $$\hat{\Lambda}(t) = \sum_{t_i \leq t} \frac{d_i}{R_i}.$$

  There exists a vast literature studying asymptotic results for these two estimators and it has bee proven that they are asymptotically equivalent. Other non-parametric estimators such as the Breslow estimator of the cumulative hazard (Breslow, 1972, 1974) and the Harrington and Flemming estimator of the survival function are respectively derived from the Kaplan-Meier and the Nelson-Aalen estimators.

  The main advantage of non-parametric estimators is their flexibility. However they do not allow for incorporating covariates.

- Semi-parametric models or proportional hazard models address the problem of incorporating covariates. The most commonly used one is called the Cox model. The hazard function is decomposed into a baseline hazard λ_0 shared across patients and a relative risk modeling the effect of covariates:

  $$\lambda(t|X_i) = \lambda_0(t) \exp\left(X_i\theta\right)$$

  where $X_i \in \mathbb{R}^{1\times p}$ is the vector of covariates for patient i and $\theta \in \mathbb{R}^{p\times 1}$ a parameter which does not depend on t. The baseline hazard λ_0 is estimated

non-parametrically. The following two assumptions must be satisfied: 1) Proportional hazard, i.e. hazard rate ratio between two patients must be independent of t, 2) Linearity between covariates and log hazard. The Cox model can be generalized to time-dependent covariates.

An extension of proportional-hazard models called frailty models incorporate a random effect which has a multiplicative effect on the hazard (Hougaard, 1995; Wienke, 2010). They are used for considering the influence of unobserved covariates and allow for taking into account dependencies between time-to-event data in subgroups, for instance common genetic influences between family members.

- Parametric models assume that the survival function follows a parametric distribution. The most commonly encountered distributions are the exponential, piecewise exponential, Weibull, Gamma, Gompertz-Makeham, generalized Weibull distributions among others. Such models allow for adding covariates. An exponential survival function is associated with a constant hazard rate (see Equation 2.1).

In this thesis we will work with parametric models and in particular with piecewise constant hazard functions.

2.2 Introduction to multi-state models

Multi-state models are applied in a wide range of domains and in particular in biomedical area as they allow for modeling the evolution of a stochastic process. They are for instance particularly adapted for studying longitudinal data such as repeated measures of a biomarker or the evolution of the health history of a patient via complex dynamics between various states. They are often used for studying the progression of a disease from an healthy state to diseased states through various severity stages or the evolution of a patient history through several diseases. A change of state is called a transition or an event and is represented by an arrow. A reversible transition is represented by two arrows in opposite directions. A state is said to be absorbant if it precludes any other transition (for instance “dead” is an absorbant state when studying the health history of a patient). Non-absorbant states are called transient states. There exists a vast literature on multi-state models. This section is based on Andersen (1988); Hougaard (1999); Commenges (1999); Saint Pierre (2005); Meira-Machado et al. (2009); Hougaard (2012).

The simplest multi-state model is composed of two states and one transition (Figure 2.1). In a competing risk model (Figure 2.2) several states are in competition and a patient can move from State 1 to one and only one state. In the L-progressive model (Figure 2.3), the patient transit through subsequent states (for instance, severity stages of a disease). The illness-death model (Figure 2.4) is composed of three states with two transient states (“healthy" and “diseased") and one absorbant state (“dead") with competing risks between disease and death. Note that the illness-death model contains two different paths from state “healthy to state “dead”. One

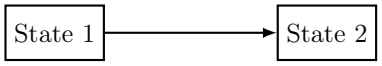


Figure 2.1: Two-state model

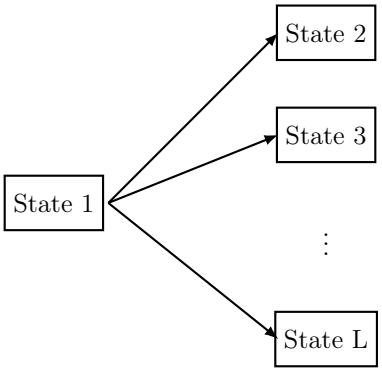


Figure 2.2: Competing risk model

may add an extra arrow from state "diseased" to state "healthy" when considering recovery. Multi-state models allow for a variety of patient histories by combining the aforementioned models for more complex histories.

We denote by $\mathcal{Z}$, the state space of the model and by $\{Z(t), t \geq 0\}$, a continuous time stochastic process where $Z(t)$ denotes the state occupied at time t. The transition intensity (or hazard) from state $k \in \mathcal{Z}$ to state $\ell \in \mathcal{Z}$ represent the instantaneous risk of moving from state k to state ℓ:

$$\lambda_{k\ell}(t, \mathcal{F}_t) \stackrel{\text{def}}{=} \lim_{\delta t \to 0} \frac{\mathcal{P}(Z(t+\delta t) = \ell | Z(t) = k, \mathcal{F}_t)}{\delta t} \tag{2.2}$$

where $\mathcal{F}_t$ is the past history at time t. Time scale may also include the calendar time in particular in population studies regarding a disease whose incidence rate varies through time (for instance HIV). Calendar time will not be considered in the framework of this thesis.

Markov property refers to the memoryless property of the process, i.e. the current state resumes information on previous states. A model is said to be Markov when all transition intensities are independent of $\mathcal{F}_t$, i.e. $\lambda_{k\ell}(t, \mathcal{F}_t) = \lambda_{k\ell}(t)$. In semi-Markov models, transition intensities are functions of the sojourn duration in the current state denoted d, i.e. $\lambda_{k\ell}(t, \mathcal{F}_t) = \lambda_{k\ell}(t, d)$. A model is said to be homogeneous when all transition intensities are not functions of t and non-homogeneous otherwise. Both notions can be combined such that we can have homogeneous Markov models (constant transition intensities), non-homogeneous Markov models ($\lambda_{k\ell}$ functions of t), homogeneous semi-Markov models ($\lambda_{k\ell}$ functions of d) or non-homogeneous

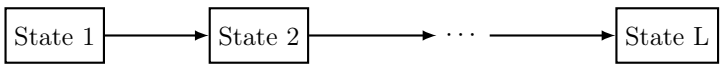


Figure 2.3: L-progressive model

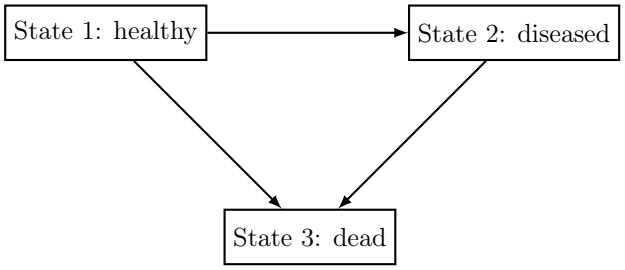


Figure 2.4: Illness-death model

semi-Markov models ($\lambda_{k\ell}$ functions of t and d). A stochastic process under Markov property is a Markov chain. Data consist in observations at given times. Observation errors can be considered with hidden Markov models. Multi-state models under Markov property have been applied in a wide range of medical fields, for instance for the analysis of biological markers in the study of the progression of cancer (Kay, 1986; Chen et al., 1996), diabetes (Marshall and Jones, 1995) or HIV (Gentleman et al., 1994; Guihenneuc-Jouyaux et al., 2000), in the study of health evolution after lung transplantation (Jackson and Sharples, 2002), the progression through staged severity of abdominal aortic aneurysms (Jackson et al., 2003). In some medical situations, semi-Markov models are needed for more realistic applications but involve more complex computations. They have been applied for instance for studying HIV progression (Foucher et al., 2005). Markov property is commonly assumed and we focus in this thesis on Markov models and hidden Markov models. We denote by

$$Q(t) = \left(\lambda_{k\ell}(t)\right)_{k,\ell\in\mathcal{Z}}$$

the $|\mathcal{Z}| \times |\mathcal{Z}|$ matrix of (possibly null) transition intensities at time t whose rows sum to zero such that, for all $k \in \mathcal{Z}$, $\lambda_{kk}(t) = -\sum_{\ell\neq k} \lambda_{k\ell}(t)$.

One of the key question in the analysis a stochastic process with a multi-state model is the computation of transition probabilities defined, under Markov property, by

$$p_{k\ell}(s,t) = \mathcal{P}(Z(t) = \ell | Z(s) = k), \tag{2.3}$$

the probability of occupying state ℓ at time t conditional on occupying state k at time s. The transition probability matrix is given by

$$P(s,t) = \left(p_{k\ell}(s,t)\right).$$

Transition probabilities are given by solving Kolmogorov forward differential equations. For homogeneous models (constant transition intensities), $P(s,t)$ is given by the exponential of the scaled matrix of transition intensities:

$$P(s,t) = \exp\left(Q \times (t-s)\right).$$

Transition probability matrices are analytically calculated and implemented in the `msm` package available at `https://cran.r-project.org/web/packages/msm/index.`

`html` (Jackson et al., 2011; Jackson and Jackson, 2019) for a selection of most commonly encountered time-homogeneous models. For other models, $P(s,t)$ is computed from matrix exponential with formal calculation. The `msm` package adresses the problem of fitting Markov models to panel data where observations are made at a finite series of times. They authors propose extensions to hidden Markov models (HMM) when observations are made through error-prone markers. The matrix exponential is given by the power series $\exp(Q) = \sum_{k=0}^{\infty} Q^k/k!$. The authors of the `msm` package recall the difficulty to reliably calculate and cite (Moler and Van Loan, 2003), the advantages of calculating the analytical expression of $P(s,t)$ in terms of Q for simple models and the usefulness of the Mathematica software for obtaining these expressions. For non-homogeneous models, $P(s,t)$ cannot be calculated in close *formulae* except if transition intensities are piecewise constant.

The contribution of each individual to the likelihood of Q, denoted $\mathcal{L}(Q)$, is given by the product of transition probabilities at the observed times for the individual. Individual contributions are assumed to be independent and multiplied for obtaining $\mathcal{L}(Q)$. One can use standard optimization methods for maximizing $\mathcal{L}(Q)$. In case of piecewise constant transition intensities, the likelihood is summed over the unobserved states at breakpoints. Questions arising in chapters involving time-to-event data in the thesis are related to computing transition probabilities and, in general, probabilities of particular histories as well as posterior probabilities of future histories with fixed parameters. No parameter estimation was performed in the framework of survival analysis du to insufficient data. Therefore, in this chapter, a particular emphasis is put on the computation of transition probabilities with given transition intensities.

2.3 Discretized piecewise constant Markov model

In order to avoid the analytical or formal calculation of matrix exponential, we propose in this section to introduce an alternative method for piecewise constant Markov models based on a time discretization for computing approximate transition probabilities. Whereas previous sections where introductions, this section constitue a contribution in the field. The model presented thereafter is adaptable to any number of states (with impact on computational time complexity detailed below) and any transition structure.

Method

The model is based on a time discretization and is detailed below in the framework of constant transition intensities, the extension to piecewise constant models being straightforward. We denote by $\mathcal{Z} = \{1, \ldots, L\}$, the state space of the model and by $\alpha_{k\ell}$ the transition intensity from state $k \in \mathcal{Z}$ to state $\ell \in \mathcal{Z}$. Using notation introduced in the previous paragraph, the $L \times L$ matrix of transition intensities $Q = (\alpha_{k\ell})$ is such that $\alpha_{kk} = -\sum_{\ell \neq k} \alpha_{k\ell}$ and we denote by $p_{k\ell}(s,t) = \mathcal{P}(Z(t) = \ell | Z(s) = k)$ the probability of occupying state ℓ at time t conditional on occupying state k at time s. The method is based on a time discretization over $(i \times \Delta)_{i=1,\ldots,n}$

where Δ is a chosen (small) step of time (for instance, a twelfth of a year for a study of lifetime events) and n is determined such that n/Δ is a chosen maximal age. The following strong assumption is made:

- We assume that a maximum of one transition can occur in any time interval $]i\Delta - \Delta, i\Delta]$.

Let $Z_i = Z(i\Delta) \in \mathcal{Z}$ be the state occupied at time $i\Delta$, we consider the Markov chain $Z = (Z_i)_{i=1,\dots,n}$ with transition probabilities $\pi_i(k,\ell) = \mathbb{P}(Z_i = \ell | Z_{i-1} = k)$ defined, for all $i \in \{1,\dots,n\}$ and $k, \ell \in \mathcal{Z}$, $\ell \neq k$, by

$$\pi_i(k,k) = \exp(\alpha_{kk}\Delta) \qquad \text{and} \qquad \pi_i(k,\ell) = (1 - \pi_i(k,k))(-\alpha_{k\ell}/\alpha_{kk}) \tag{2.4}$$

Under the hypothesis of a maximum of one event in time interval $]i\Delta - \Delta, i\Delta]$, the expression of $\pi_i(k,k)$ is straightforward as the sojourn time in state k follows an exponential distribution of rate $-\alpha_{kk}$. The expression of $\pi_i(k,\ell)$, $\ell \neq k$, is an analytical expression of the integration between $i\Delta - \Delta$ and $i\Delta$ of the density of the time of entrance in state $\ell \neq k$ conditional on occupying state k at time $i\Delta - \Delta$ given, for $t \in]i\Delta - \Delta, i\Delta]$, by $\exp(\alpha_{kk} \times [t - (i\Delta - \Delta)]) \times \alpha_{k\ell}$.

The classical forward-backward algorithm (see Rabiner, 1989; Durbin et al., 1998; Cappé et al., 2005) is used to compute any approximate transition probability in linear time complexity. We briefly recall its implementation using notation and definitions seen in Chapter 1 and, in the particular case of a Markov chain, in Section 1.7.2. We denote by ev an evidence for Z (see Section 1.1.1.2) such that $\text{ev} = \cap_{i \in E}\{Z_i \in \mathcal{Z}_i^* \subset \mathcal{Z}\}$ where $E \subseteq \{1,\dots,n\}$ and, for all $i \notin E$, let $\mathcal{Z}_i^* = \mathcal{Z}$. We introduce the potentials

$$\phi_i(Z_{i-1} = k, Z_i = \ell) = \pi_i(k,\ell) \times \eta_i(k,\ell)$$

where $Z_0 = 1$ by convention and $\eta_i : \mathcal{Z}^2 \mapsto \{0,1\}, \eta_i(k,\ell) = \mathbb{1}_{\{(k,\ell) \in \mathcal{Z}_{i-1}^* \times \mathcal{Z}_i^*\}}$. Note that $\mathbb{P}(Z_1,\dots,Z_n,\text{ev}) \propto \prod_{i=1}^n \phi_i(Z_{i-1}, Z_i)$. We define respectively the forward and backward messages to be, for all $i = 1,\dots,n$,

$$F_i(Z_i) \stackrel{\text{def}}{=} \mathbb{P}(Z_1 \in \mathcal{Z}_1^*, \dots, Z_i \in \mathcal{Z}_i^*) \quad \text{and} \quad B_i(Z_i) \stackrel{\text{def}}{=} \mathbb{P}(Z_{i+1} \in \mathcal{Z}_{i+1}^*, \dots, Z_n \in \mathcal{Z}_n^* | Z_i).$$

Forward messages can be recursively computed with the forward pass of the forward-backward algorithm in $\mathcal{O}\left(nL^2\right)$ time complexity with initialization $F_1(1,k) = \pi_1(1,k)$ and for all $i = 2,\dots,n$ and $k, \ell \in \mathcal{Z}$, $F_i(\ell) = \sum_{k \in \mathcal{Z}} F_{i-1}(k)\phi_i(k,\ell)$. Similarly backward messages are computed in $\mathcal{O}\left(nL^2\right)$ time complexity, using a backward pass with initialization, for all $k \in \mathcal{Z}$, $B_n(k) = 1$ and for all $i = n,\dots,2$ and $k, \ell \in \mathcal{Z}$, $B_{i-1}(k) = \sum_{\ell \in \mathcal{Z}} \phi_i(k,\ell)B_i(\ell)$. Finally we have, for all $i \in \{1,\dots,n\}$ (see Theorem 2),

$$\mathbb{P}(Z_i, \text{ev}) = F_i(Z_i)B_i(Z_i). \tag{2.5}$$

Hence one can compute an approximate probability of a configuration of interest $\mathbb{P}(\text{ev})$ as well as an approximate posterior probability $\mathbb{P}(Z_i|\text{ev})$ in $\mathcal{O}\left(nL^2\right)$. Note that for the probability of a configuration, the forward pass is sufficient as one can compute $\mathbb{P}(\text{ev}) = \sum_{k \in \mathcal{Z}} F_n(k)$ choosing index $i = n$ in Equation (2.5).

Remark 1: A configuration of interest can be a variety of quantities including any desired transition probability as defined in the multi-state survival framework (see Equation 2.3) by setting ev appropriately. In order avoid any confusion between a transition probability defined as such and a transition probability defined in HMMs, we will sometimes refer to a state occupancy rather than a transition probability if necessary.

Remark 2: The extension to piecewise constant transition intensities with cuts $c = (c_1, \ldots, c_N)$ is straightforward with a choice of Δ such that $c \subseteq (i \times \Delta)_{i=1,\ldots,n}$.

The main advantages of the discretized HMM is its easy implementation, it avoids the computation of exponential matrices and it can be adapted to any number of states and transition structure. It's main drawback is the assumption of a maximum of one event per interval of size Δ hence we solely compute approximate transition probabilities. The choice of Δ is made as a compromise between augmentation of time complexity and precision according model structure and magnitude of transition intensities.

Results

We denote by dMM the discretized Markov model detailed in the previous paragraph and we propose in this section an illustration of quantities computed with dMM over two simple examples: the illness-death model with no recovery represented in Figure 2.4 and constant transition intensities as well as a simpler version of the model that will be used in Chapter 8 with piecewise constant transition intensities.

Let us start this result section with a brief comparison of some transition probabilities computed with dMM and analytically in the illness-death model represented in Figure 2.4 and constant transition intensities. The matrix of transition intensities is given by

$$Q = \begin{pmatrix} -(\alpha_{12} + \alpha_{13}) & \alpha_{12} & \alpha_{13} \\ 0 & -\alpha_{23} & \alpha_{23} \\ 0 & 0 & 0 \end{pmatrix}$$

where, for $k, \ell \in \{1, 2, 3\}$, $\alpha_{k\ell}$ denotes the transition intensity from state k to state ℓ. The analytical expression of each element of the matrix of transition probabilities $P(s, t) = (p_{k\ell}(s, t))$ in that particular framework is recalled in (Jackson and Jackson,

2019) and given thereafter:

$$
\begin{aligned}
p_{11}(s,t) &= e^{-(\alpha_{12}+\alpha_{13})(t-s)} \\
p_{12}(s,t) &= \begin{cases} \frac{\alpha_{12}}{\alpha_{12}+\alpha_{13}-\alpha_{23}}\left(e^{-\alpha_{23}(t-s)} - e^{-(\alpha_{12}+\alpha_{13})(t-s)}\right) & \text{if } \alpha_{12}+\alpha_{13} \neq \alpha_{23} \\ \alpha_{12}(t-s)e^{-(\alpha_{12}+\alpha_{13})(t-s)} & \text{if } \alpha_{12}+\alpha_{13} = \alpha_{23} \end{cases} \\
p_{13}(s,t) &= \begin{cases} 1 - e^{-(\alpha_{12}+\alpha_{13})(t-s)} - \frac{\alpha_{12}}{\alpha_{12}+\alpha_{13}-\alpha_{23}}\left(e^{-\alpha_{23}(t-s)} - e^{-(\alpha_{12}+\alpha_{13})(t-s)}\right) & \text{if } \alpha_{12}+\alpha_{13} \neq \alpha_{23} \\ \left(-1 + e^{(\alpha_{12}+\alpha_{13})(t-s)} - \alpha_{12}(t-s)\right)e^{-(\alpha_{12}+\alpha_{13})(t-s)} & \text{if } \alpha_{12}+\alpha_{13} = \alpha_{23} \end{cases} \\
p_{21}(s,t) &= 0 \\
p_{22}(s,t) &= e^{-\alpha_{23}(t-s)} \\
p_{23}(s,t) &= 1 - e^{-\alpha_{23}(t-s)} \\
p_{31}(s,t) &= 0 \\
p_{32}(s,t) &= 0 \\
p_{33}(s,t) &= 0.
\end{aligned} \tag{2.6}
$$

We (arbitrarily) assume that $\alpha = (\alpha_{12}, \alpha_{13}, \alpha_{23}) = (1/120, 1/100, 1/50)$. For $k \in \{1,2,3\}$, $\mathbb{P}(Z(t) = k) \approx p_{1,k}(0,t)$ is computed at time $t = 20$, $t = 50$ and $t = 80$ with dMM (Equation 2.5) with $Z_0 = 1$ and evidence $\text{ev} = \{Z_{t/\Delta} = k\}$ using various chosen steps of time Δ ($\Delta = 1$, $\Delta = 0.1$, $\Delta = 0.01$). Secondly $p_{1,k}(0,t)$ is computed analytically with Equations (2.6). Computed quantities are compared in terms of their ratio and their difference. Results are reported in Table 2.1. Note that under the assumption of a maximum of one event in any time interval $]i\Delta - \Delta, i\Delta]$, no error is made when computing transition probabilities of the form $p_{11}(s,t)$ using dMM (the individual remains in state 1) and therefore these results are excluded from Table 2.1. Recalling that $p_{12}(0,t) = \int_0^t \exp(-\int_0^u (\alpha_{12}+\alpha_{13})\mathrm{d}v)\alpha_{12}\exp(-\int_u^t \alpha_{23}\mathrm{d}w)\mathrm{d}u$, the upper bond of ratios of results associated with $p_{12}(0,t)$ lays in the last term of this expression and is given, for any $t \geq 0$, by $\exp(\alpha_{23}\Delta)$. Indeed, for any $u \geq 0$, death is precluded in a time interval of length strictly below Δ. As expected, transition probabilities of the form $p_{12}(0,t)$ (respectively $p_{13}(0,t)$) are overestimated (respectively underestimated) by dMM and ratios of results associated with $p_{12}(0,t)$ are constant in t and strictly lower than $\exp(\alpha_{23}\Delta)$ respectively equal to 1.0202, 1.0020 and 1.0002 for $\Delta = 1$, $\Delta = 10$ and $\Delta = 100$.

In order to introduce the model developed in the last part of the thesis, we propose in the second part of this section an overview of computed quantities of particular interest in medical genetics with a simpler version of the model used in Chapter 8 and we leave the genetic part aside. We consider two diseases A and B and we are interested in computing the probability of a personal history of disease and future disease risks under the assumption of a maximum of two diagnoses per individual up to age 80. The associated multi-state model is represented in Figure 2.5 where State 1 or U stands for healthy, State 2 or A (respectively State 3 or B) stands for diagnosed with disease A (respectively with disease B), State 4 or AA (respectively State 5 or AB) stands for diagnosed with A (respectively B) after A and State 6 or BA (respectively State 7 or BB) stands for diagnosed with A (respectively B) after B.

	$p_{12}(0,20)$	$p_{12}(0,50)$	$p_{12}(0,80)$	$p_{13}(0,20)$	$p_{13}(0,50)$	$p_{13}(0,80)$
Analytically	0.1136029	0.1598511	0.1439833	0.1933565	0.4402993	0.6253235
$\Delta = 1$						
dMM	0.1147462	0.1614598	0.1454324	0.1922132	0.4386905	0.6238744
Ratio	1.0100642	1.0100642	1.0100642	0.994087	0.9963462	0.9976827
Error (diff.)	0.0011433	0.0016088	0.0014491	-0.0011433	-0.0016088	-0.0014491
$\Delta = 0.1$						
dMM	0.1137165	0.1600110	0.1441274	0.1932428	0.4401393	0.6251794
Ratio	1.0010006	1.0010006	1.0010006	0.9994121	0.9996367	0.9997696
Error (diff.)	0.0001137	0.0001600	0.0001441	-0.0001137	-0.0001600	-0.0001441
$\Delta = 0.01$						
dMM	0.1136142	0.1598671	0.1439977	0.1933451	0.4402833	0.6253091
Ratio	1.0001000	1.0001000	1.0001000	0.9999412	0.9999637	0.999977
Error (diff.)	0.0000114	0.0000160	0.0000144	-0.0000114	-0.0000160	-0.0000144

Table 2.1: Transition probabilities computed analytically (first line) and with dMM for $\alpha = (1/120, 1/100, 1/50)$ and Δ varying. Associated errors are given in terms of ratio and differences of results computed respectively with dMM and analytically. Grayed values are non-informative.

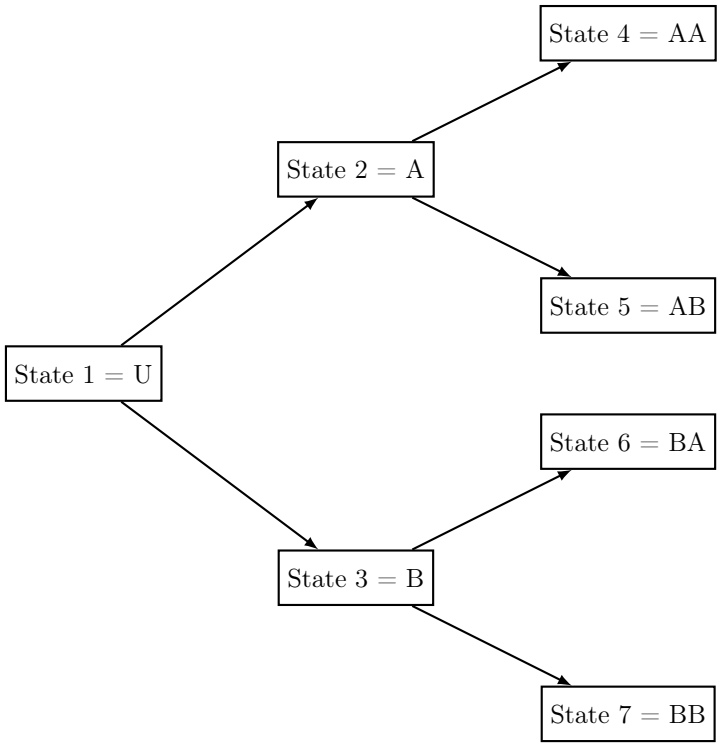


Figure 2.5: Multi-state model with two diseases, A and B, and a maximum of two diagnoses.

For all $k, \ell \in \mathcal{Y} = \{1, \ldots, 7\}$, we denote by $\lambda_{k\ell}$ the transition intensity from state k to state ℓ and we assume that transition intensities are piecewise constant with commun cuts $c = (c_1 = 20, c_2 = 40, c_3 = 60)$. Transition intensities are defined for all $k, \ell \in \mathcal{Y}$ and $t \geq 0$, by

$$\lambda_{k\ell}(t) = \mathbb{1}_{t \in]c_{j-1}, c_j]} \alpha_{k\ell,j}$$

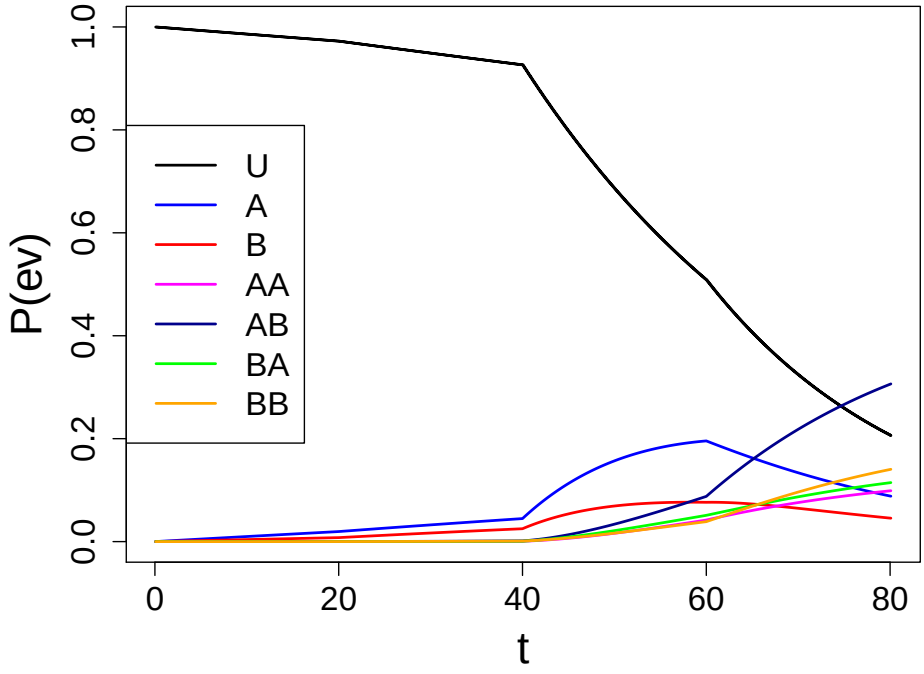
where $c_0 = 0$, $c_4 = \infty$ by convention and $\alpha_{k\ell,j}$ is the hazard rate from state k to state ℓ in time interval $]c_{j-1}, c_j]$. Vectors of hazard rates denoted $\alpha_{k\ell} = (\alpha_{k\ell,1}, \ldots, \alpha_{k\ell,4})$ are arbitrarily chosen and we assume that $\alpha_{12} = (1/1000, 1/700, 1/50, 1/40)$, $\alpha_{13} = (4/10000, 1/1000, 1/100, 1/50)$, $\alpha_{24} = \alpha_{12} \circ (1/1.8, 1/1.5, 1/1.4, 1/1.2)$, $\alpha_{25} = \alpha_{13} \circ (1, 2, 3, 4)$, $\alpha_{36} = \alpha_{12} \circ (1, 1.5, 2, 2)$ and $\alpha_{37} = \alpha_{13} \circ (2, 3, 3, 4)$ where $\circ$ is the Hadamard product. We choose $\Delta = 1/12$ and a maximal age 80, hence n is set to $n = 80/\Delta$.

Available time-to-event data in medical genetics and genetic counseling are fairly often prone to uncertainty as a patient reports some histories of disease of family members up the highest degree relative he/she can remember. Some data may include uncertain status (for instance, gynecological cancer rather than endometrial, ovarian, etc. cancer) or a time interval rather than a precise age of diagnosis or last news. We selected in this section various evidences in order to cover most commonly encountered data in medical genetics and propose an overview of computed probabilities of state occupancy and posterior probabilities of interest using dMM. Some (non-exhaustive) comments are added to show the interest of the model and the qualitative coherence of the results.

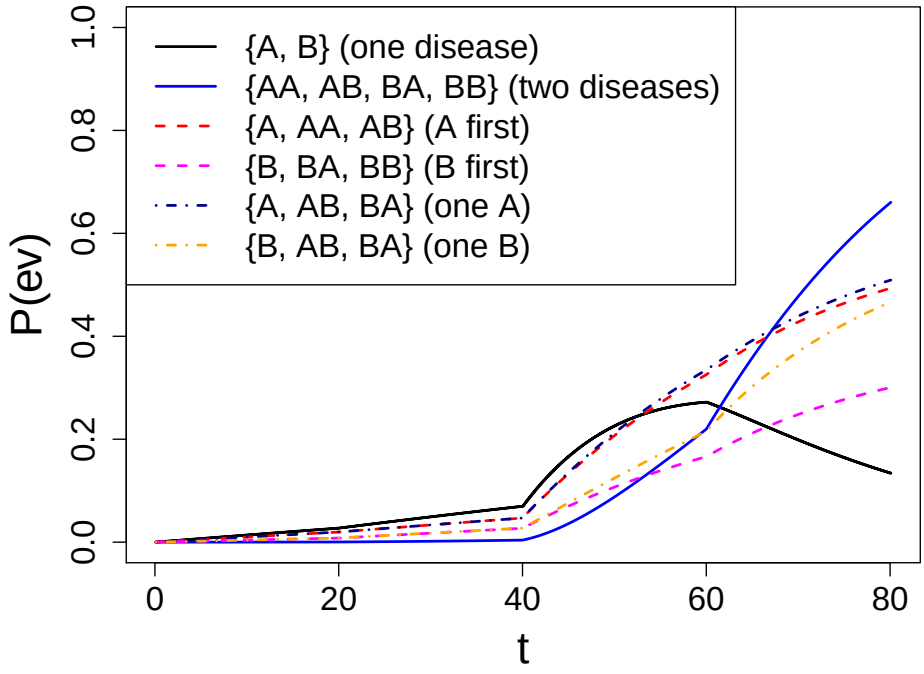
Remark: Note that one could use the model represented in Figure 2.7 of lower state space, hence a lower time complexity, for computing the probability of a configuration of interest $\mathbb{P}(\text{ev})$. For instance, one can compute the probability of occupying state $\cdot$B after A by defining ev over couples (Z_{i-1}, Z_i) such that $\eta_i(k, \ell) = 0$ for all $i \in \{1, \ldots, n\}$ and all couples $(k, \ell) \notin \{(1, 1), (1, 2), (2, 2), (2, 5), (5, 5)\}$. However such model remain inappropriate for computing posterior probabilities of the form $\mathbb{P}(Z_i|\text{ev})$. For the sake of simplicity, all results below are presented using the model in Figure 2.5.

In order to ease the reading, a state will either be denoted by its associated value in the set $\mathcal{Y} = \{1, \ldots, 7\}$ or its associated letter(s) in the set $\mathcal{Z} = \{$U, A, B, AA, AB, BA, BB$\}$, whichever best sustains the fluidness of the manuscript.

Let us firstly propose in Figure 2.6 a graphical representation of the probability of occupying state $k \in \mathcal{Y}$, i.e. $\mathbb{P}(Z_{t/\Delta} = k|Z_0 = 1) \approx p_{1k}(0, t)$ (Figure 2.6a) and the probability of occupying any state $k \in K \subset \mathcal{Y}$ for chosen sets K, i.e. $\mathbb{P}(Z_{t/\Delta} \in K \subset \mathcal{Y}|Z_0 = 1) \approx \sum_{k \in K} p_{1k}(0, t)$ (Figure 2.6b) computed at age t varying from 0 to 80 by steps of size Δ. Note a greater probability of being diagnosed with a single disease A (state A) than a single disease B (state B) at any age explained by transition intensities $\lambda_{12}(t) \geq \lambda_{13}(t)$ for all $t \geq 0$. Moreover the steep decrease of the probability of occupying state A after 60 is consistent with the steep increase of the probability of occupying state AB after 60 and explained by high values of $\alpha_{25,4}$. A similar remark can be made when comparing the probability of occupying state B and state BB. In Figure 2.6b we can see that the probability of being diagnosed with two diseases ($K = \{\text{AA}, \text{AB}, \text{BA}, \text{BB}\}$) overpasses the one of having developed one disease ($K = \{\text{A}, \text{B}\}$) from a fairly young age (62 years old) partly explained by greater transition



(a) Probability $\mathbb{P}(Z_{t/\Delta} = k \in \mathcal{Z})$ of occupying a state $k \in \mathcal{Z}$ at age t.



(b) Probability $\mathbb{P}(Z_{t/\Delta} \in K \subset \mathcal{Z})$ of occupying any state $k \in K \subset \mathcal{Z}$ at age t

Figure 2.6: Graphical representation of probabilities of state occupancy at age t.

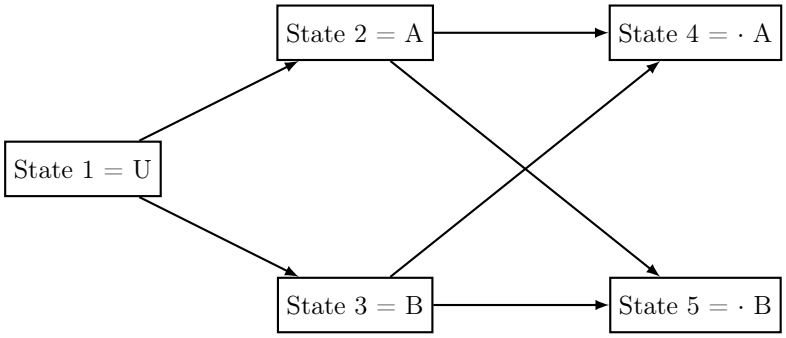


Figure 2.7: Another example of a multi-state model with two diseases, A and B, and a maximum of two diagnoses.

intensities from diseased states when compared to transition intensities from the healthy state, except λ_{24}. Let us finally highlight a higher probability of being diagnosed with A before B ($K = \{\text{A}, \text{AA}, \text{AB}\}$) than B before A ($K = \{\text{B}, \text{BA}, \text{BB}\}$) mostly explained by transition intensities $\lambda_{12}(t) \geq \lambda_{13}(t)$ for all $t \geq 0$.

The posterior probability of state occupancy conditional on being diagnosed with A before B, both diseases before age 70 is represented in Figure 2.8. States associated with constantly null posterior probabilities are not reported. Note the increasing posterior probability of occupying state AB until overpassing the one of occupying state A and reaching one from age 70. The derivative of the that probability increases at each cut $c_1 = 20$, $c_2 = 40$, $c_3 = 60$ which is consistent with increasing values of α_{25}.

Figure 2.9 represents computed posterior probabilities of state occupancy conditional on a first diagnosed disease being A at age 50 (Figure 2.9a) (respectively conditional on a diagnosis of disease A at age 50 (Figure 2.9b)). Note in Figure 2.9a a constant posterior probability equal one of occupying state "U" for all $t < 50$ and a probability of occupying state "A" at age 50 equal one. As the individual can move from state A to state AA or AB, posterior probabilities of occupying these two latter states increase after age 50 at greater speed for AB du to a greater value of $\alpha_{25,j}$ than $\alpha_{24,j}$ for $j \in \{3, 4\}$ (respectively in time interval $]40, 60]$ and $]60, +\infty]$). In Figure 2.9b, the individual may have encountered A or B before being diagnosed with A at age 50, leading to non-null posterior probabilities for all states before age 50 except BB (the individual must have encountered A at age 50 leading to a null posterior probability of occupying state BB at any age). In particular for instance the posterior probability of occupying state B is not null before age 50 and becomes null after age 50 as the individual must occupy either state A, AB or BA after age 50. Note that the individual may have encountered B before begin diagnosed with A at age 50, hence occupy state BA at age 50. BA being an absorbant state, the posterior probability of occupying that state is constant after age 50. Unlike the posterior probability of occupying state AA, the posterior probability of occupying state AB is of course null before age 50 as A must have been diagnosed at age 50 but increases at greater speed after age 50 for the same reasons as the ones seen with Figure 2.9a.

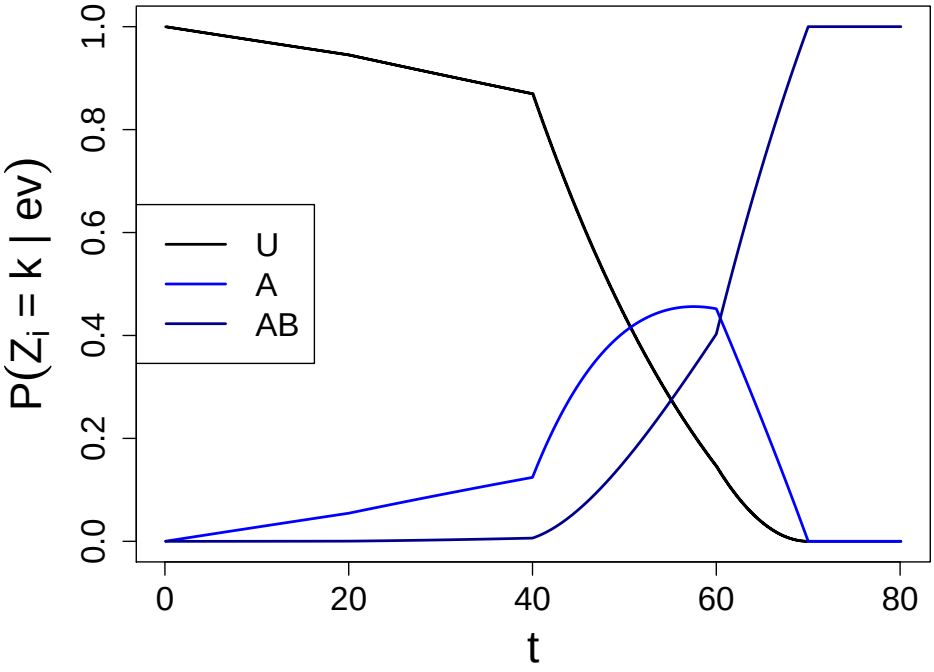
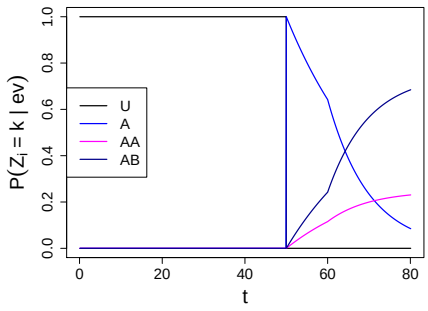
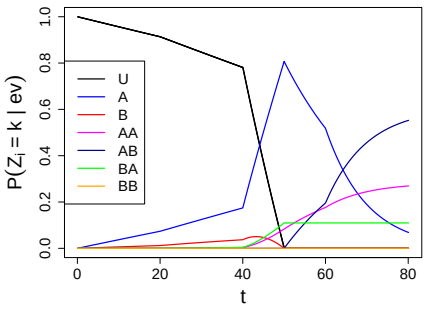


Figure 2.8: Marginal posterior probabilities of state occupancy conditional on being diagnosed with A before B, both diseases before age 70.

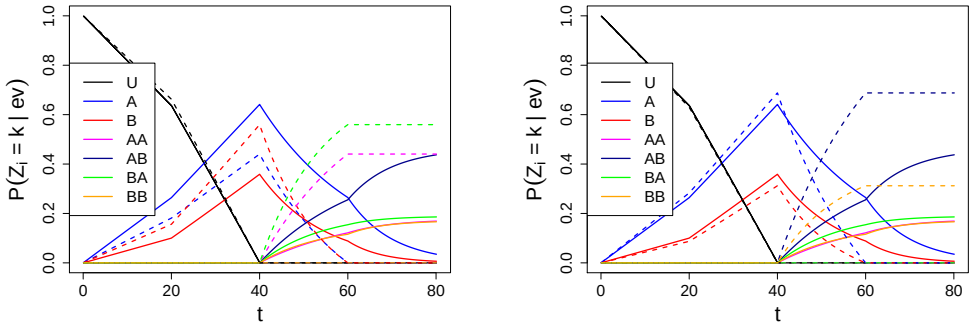


(a) Posterior probabilities conditional on first diagnosed disease being A at age 50.



(b) Posterior probabilities conditional on a diagnosed disease A (not necessarily the first) at age 50.

Figure 2.9: Marginal posterior probabilities of state occupancy respectively conditional on first diagnosed disease being A at age 50 (on the left) and conditional on a diagnosed disease A, not necessarily the first, at age 50 (on the right).



(a) Second diagnosis of disease A.

(b) Second diagnosis of disease B.

Figure 2.10: Marginal posterior probabilities of state occupancy conditional on a first diagnosed disease A or B before age 40 (plain lines) and an additional diagnosis at age 60 (dashed lines) respectively of disease A (on the left) and disease B (on the right).

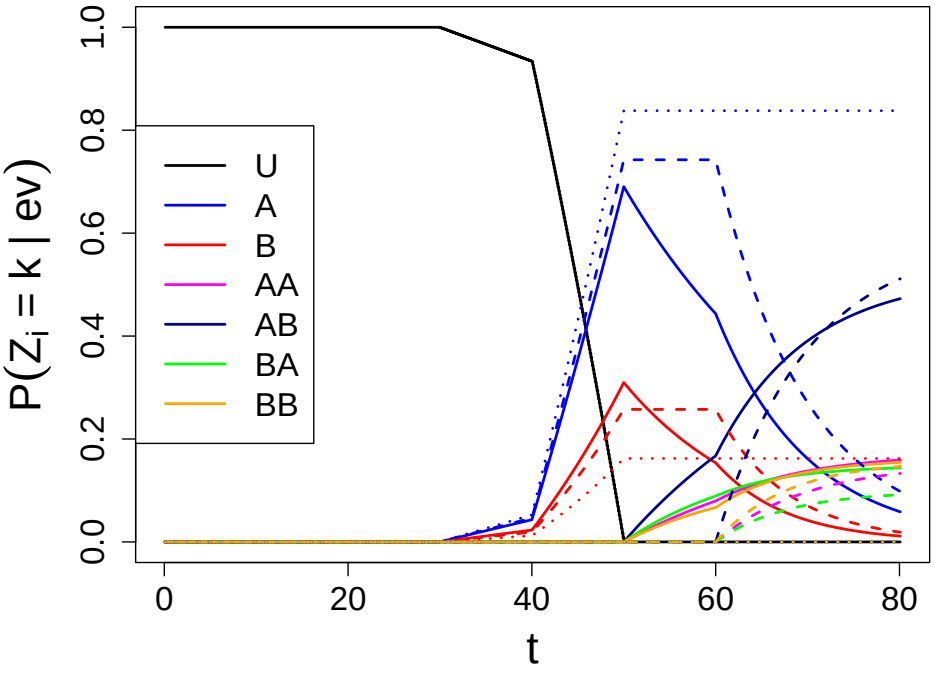


Figure 2.11: Marginal posterior probabilities of state occupancy conditional on being diagnosed with one and only one first disease A or B in time interval $]30, 50]$ (plain lines) and no other disease up to age 60 (dashed lines) (respectively 80, dotted lines).

In Figure 2.10, we propose a graphical representation of posterior probabilities of state occupancy conditional on a first diagnosed disease A or B before age 40 (plain lines) and an additional diagnosis at age 60 (dashed lines) respectively of disease A (Figure 2.10a) and disease B (Figure 2.10b). Note the null posterior probabilities of occupying any state $k \notin \{\text{U, A, B}\}$ before age 40 which remains null for $k \in \{\text{AB, BB}\}$ (respectively $k \in \{\text{AA, BA}\}$) when conditioning respectively on second disease being A at 60 (dashed lines in Figure 2.10a) and second disease being B at age 60 (dashed lines in Figure 2.10b) which is consistent with the respective evidence. Among several other remarks, we note that the additional conditioning on second diagnosis of disease A (respectively B) at age 60 decreases (respectively increases) the posterior probability of first diagnosis with A partly explained and consistent with values taken by transition intensities in time interval $]40, 60]$. Indeed we have $\alpha_{24,3} < \alpha_{36,3}$ and $\alpha_{25,3} = \alpha_{37,3}$. Therefore an individual diagnosed with a second disease being A at age 60 is more likely to occupy state B rather than A just before 60 (Figure 2.10a). On the contrary, an individual not diagnosed with A nor B in time interval $]40, 60]$ is more likely to be in state A rather than B in time interval $]40, 60]$ and the next transition associated with a diagnosis of B at age 60 is as likely for individuals occupying state A or B just before 60 (Figure 2.10b).

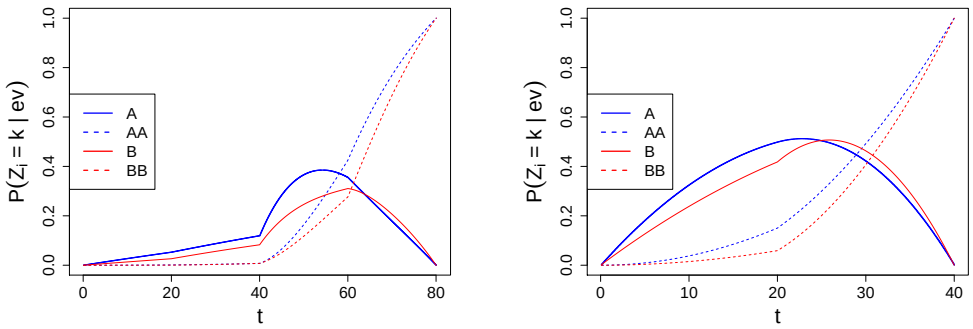
Figure 2.11 pictures computed posterior probabilities of state occupancy conditional on being diagnosed with a single first disease A or B in time interval $]30, 50]$ (ev_1, plain lines) and being free of other disease respectively up to age 60 (ev_2, dashed lines) and 80 (ev_3, dotted lines). Note some straightforward first remarks such that the posterior probability of occupying state “U" being 1 before 30 and the null posterior probabilities of occupying any state $k \in \{\text{AA, AB, BA, BB}\}$ at any age below 30, 60 and 80 conditional respectively on ev_1, ev_2 and ev_3. Moreover we have, for all $t \in]30, 50]$, $\mathbb{P}(Z_{t/\Delta} = A|\text{ev}_3) > \mathbb{P}(Z_{t/\Delta} = A|\text{ev}_2) > \mathbb{P}(Z_{t/\Delta} = A|\text{ev}_1)$ (and respectively $\mathbb{P}(Z_{t/\Delta} = B|\text{ev}_3) < \mathbb{P}(Z_{t/\Delta} = B|\text{ev}_2) < \mathbb{P}(Z_{t/\Delta} = B|\text{ev}_1)$ which is consistent with hazard rates being equal or higher from state B than state A. Indeed, the latter an individual seen free of second disease after age 50, the greater the probability of occupying state A and not B at age 50.

Finally the posterior distribution of the age at first diagnosis (plain lines) and second diagnosis (dashed lines) is represented in Figure 2.12 conditional on being diagnosed respectively with two diseases A (blue lines) and two diseases B (red lines) before age 80 (Figure 2.12a) or 40 (Figure 2.12b).

This section is solely a brief overview of some computed quantities of interest in order to show the flexibility of the model and coherence of results. Extended empirical verifications need to be done and theoretical properties developed. This will be done in future work.

2.4 Conclusion

In this chapter, we firstly proposed two introductory sections composed of main definitions, functions and models used in the field of survival analysis as well as an introduction to multi-state models. Multi-state models are applied in a variety of medical fields, in particular for instance, for studying the evolution of a patient



(a) Diagnoses before age 80.

(b) Diagnoses before age 40.

Figure 2.12: Distribution of the age at first diagnosis (plain lines) and second diagnosis (dashed lines) conditional on being diagnosed respectively with two diseases A (blue lines) and two diseases B (red lines) before age 80 (on the left) or 40 (on the right)

through different disease states. In that sense a multi-state model is one of the main component of the model developed in Chapter 8 for computing risks of genetic predisposition and cancer risks in the framework of the Lynch syndrome.

We secondly proposed an alternative method for computing transition probabilities and posterior state probabilities in Markov multi-state models which is, to the best of our knowledge, novel in the field. The method is based on a hidden Markov model with a time discretization and allows one for avoiding the analytical or formal calculation of matrix exponential. However one must assume that a maximum of one transition can occur in any time interval according to the chosen time step. Therefore, the chosen discretization should ensure an acceptable compromise between algorithmic complexity and precision according to model structure and magnitude of transition intensities. We proposed a selection of computed quantities of interest over simulated datasets to show the interest of the method and coherence of the results. This method will be used in Chapter 8. Several perspectives remain ahead including extensive computations and comparisons over a variety of simulation schemes and real datasets, parameter estimation as well as the development of theoretical results. Furthermore we would like to investigate the extension of the model to non piecewise constant transition intensities.

Chapter 3

Pedigree-based models

Sommaire

3.1 Heredity and genetics

3.1.1 A brief history of heredity

We define a trait, also called a phenotype, to be an observable or measurable outcome. A trait can be dichotomous (e.g. absence or presence of a disease), categorical (e.g. hair color), continuous (e.g. blood pressure) or a time-to-event data (e.g. status regarding a disease and age at diagnosis or censoring, whichever comes first). Heredity, also called inheritance, is the transmission of traits from parents to offsprings and its history starts long before the birth of genetics. Pythagoras (∼580 - ∼495 BC) firstly suggested that moist vapors travelling through the body of male collect information and transmit it to his offsprings. Hippocrate (∼460 - 377 BC), considered to be the father of medicine, later suggested that seeds were produced by different parts of both male and females bodies to create a semence. Aristotle (384 - 322 BC) was the first to propose a complete theory of heredity in his book *"Generation of animals"* published between 330 and 322 BC. Aristotle postulates that male and female fluids must be mixed at conception. According to Aristotle, the male semence is highly purified blood and transmits the form (or information) and the female semence is impure blood and transmits the nutritive material. Preformation theory later emerged from antiquity suggesting that the semence of an individual contains miniature preformed versions of an organism, which contains semence containing miniature preformed individuals, and so on, at Russian dolls manner opening doors of centuries of disagreements between animalkulism and ovism respectively supporting a male and a female lineage.

The term pangenesis etymologically comes from *pan* ("whole") and *genesis* ("birth") or *genos* ("origin") and has been introduced by Charles Darwin (1809-1882) in his book *"The Variation of Animals and Plants under Domestication"* published in 1868, nine years after his famous theory about evolution published in 1859 in his book *"On the Origin of Species"*. Through pangenesis, Darwin proposes and demonstrates a theory in an attempt to explain the concept of *inheritance of acquired characteristics* also called *soft inheritance* or *Lamarckism*. Darwin suggests that, in response to environmental factors, each cell produces small particules (still undefined) called gemmules, traveling through the body, not necessarily via the bloodstream, and aggregating in gonades (sperm and ova) before being transmitted to the offsprings.

August Weismann (1834-1914) is very skeptical about the inheritance of acquired traits although he recognizes the difficulty to prove it wrong. Weismann introduces the concept of autonomous material suggesting two different types of cells. He recognizes the sadness and inconclusive results of his experiments on mice consisting in surgically cutting their tails and observing the size of the tail of their offsprings. However he proposes a new theory of heredity conferring germ cells to be the only cells carrying hereditary information. He details his theory in his book *"The germ Plasm: a Theory of Inheritance"* published in 1892 rendering the co-existence of pangenesis and his own theory complicated. Combining Darwin's and Weismann's theories seemed impossible until genetics found a unified explanation.

In 1853, the monk Gregor Mendel (1822-1884), considered to be the father of modern genetics, conducts several botanical experiments. Mendel grows peas in

his experimental garden in a monastery in Brno (Czech Republic) and studies seven different traits that seem to be independently inherited: seed shape, flower color, seed coat tint, pod shape, unripe pod color, flower location and plant height. Regarding for instance seed shape which was either angular or round, he observes that, when crossing round and angular peas in a first generation, he solely obtains round peas in the second generation. The angular characteristic disappears. Mendel crosses peas of the second generation and notes that the angular trait reappears in smaller proportion in the third generation. He concludes that the angular shape, though note visible, is still a constitutive material in the second generation but silenced. He names the trait that never disappears the dominant one and the trait that disappears and reappears from a generation to the next, the recessive one. Pursuing some simple statistical analyses, he publishes his results in 1866 in Mendel (1866). Despite the importance of these results, later named *Mendel's laws of inheritance*, the scientific community did not give any attention to his work for more than three decades. Although scientists were aware of Mendel's work, its foundation on a single trait / single "gene" (not named gene at that time) seemed not applicable to the apparent pattern of heredity du to a blending of many traits and "multiple genes" interactions.

Hugo de Vries (1848-1935) published his book *"Intracellular pangenesis"* in 1889 in which he further develops Darwin's pangenesis theory and introduces the notion of units of hereditary information that he named *pangene*. Unaware of Mendel's work, he conducts a series of similar experiments and rediscovers Mendel's laws of inheritance. He his better known for introducing the concept of mutation and developing the mutation theory.

When Mendel's laws are rediscovered they lack the description of material supporting heredity. Between 1902 and 1904 Walter Sutton and Theodor Boveri independently develop the chromosome theory of inheritance which was consistent with Mendel's discoveries (Sutton, 1902, 1903; Boveri, 1904). Theodor Boveri sees a correlation between Mendel's laws and his work on cytology. He proves that the information that never disappears is contained in the nucleus of cells inside filiform structures called chromosomes. Wilhelm Johannsen calls these information *genes*, in relation with de Vries's pangenesis. However the chemical and physical nature of genes is still unknown by then.

Around 1910 Thomas Morgan (1866-1945) and his colleagues start a series of experiments on Drosophila melanogaster, a species of fly, and confirm Mendel's laws of inheritance except the independent transmission of traits. Morgan observes that certain traits are more often transmitted together and give birth to the theory of genetic linkage. He suggests that genes are not randomly disseminated in the nucleus but rather see them as pearls along a thread. The first genetic cartographies prove that genes are physical entities organized in space.

Mendel's laws also lack explanations about discontinuous traits and multifactorial correlations, later explained by Ronald Fisher (1890-1962), a British statistician, in a paper published in 1919 and entitled *"The correlation between relatives on the supposition of Mendelian inheritance"* (Fisher, 1919). Fisher introduces and defines the term variance and proves that continuous variations of traits can be explained by the action of multiple discrete genes, each of them following Mendel's laws of

inheritance. Fisher's discoveries were one of the first steps towards population genetics and quantitative genetics. Fisher also later combined Mendelian genetics and Darwinian natural selection in his book entitled *"The Genetical Theory of Natural Selection"* first published in 1930 (Fisher, 1930).

At the end of the second world war, the immense technological advances open a door to molecular biology. The words transmission of information and genetic code appear. The scientific community is more encline to believe in proteins, constitutive molecules of the body. But in 1944 Oswald Avery, Colin MacLeod and Maclyn McCarty work on bacterias and remove, one by one, chemical substances in order to identify the one that transports genetic information. They isolated DNA to be that one and published their results in (Avery et al., 1944).

In 1951 James Watson (1928-) and Francis Circk (1916-2004) start working together on the structure of DNA from knowledge about its chemical composition, especially an equal proportion of nucleobases adenine and thymine and of nucleobases cytosine and guanine known as the Chargaff's rules (Tamm et al., 1953). In 1952, an X-ray diffraction image of DNA taken by Rosalind Franklin and her student Raymond Gosling in Maurice Wilkins' laboratory helped Watson and Crick discover the double helix structure of DNA, bases facing together by pairs adenine-thymine and cytosine-guanine. Watson and Crick published their results in (Watson and Crick, 1953) and received in 1962 the Nobel Prize in Physiology or Medicine. The double helix structure suggests that one strand is a pattern for the other. The DNA seems to be a code carrying information and a question raises: how is it decoded and translated?

3.1.2 Fundamentals in molecular genetics

Almost all human cells, except gonades, contain 23 pairs of chromosomes, called genome, divided into 22 numbered pairs of autosomes and one pair of sex chromosomes (either XY in males or XX in females). Two chromosomes of a same pair are said to be homologous. One of them is of paternal origin and the other one of maternal origin. Gonades contain 22 autosomes and one sex chromosome (X or Y). A cell containing chromosomes by pair is said to be diploid, otherwise, it is said to be haploid. Hence, most cells are diploid and gonades are haploid.

Each chromosome is made of two oriented antiparallel strands of *deoxyribonucleic acid* (DNA). Orientation is given by 5' and 3' ends. The backbone of DNA is composed of deoxyriboses covalently bound together by phosphate molecules. A nucleic acid called a base is attached to each deoxyribose. Four bases exist in DNA divided into two purines (Adenine (A) and Guanine (G)) and two pyrimidines (Cytosine (C) and Thymine (T)). The two strands of DNA molecules are linked together by non-covalent hydrogen bonding between opposite bases forming a double-stranded molecule of helical form. The opposite complementary base facing an adenine (respectively a cytosine) is always a thymine (respectively a guanine). Neighboring bases on the same strand are denoted with a small letter p, as phosphate, in between (e.g. CpG) and opposite complementary bases, linked by an hydrogen bond, are called base-pair and simply denoted one after the other (e.g. AT).

3.1.2.1 Gene expression

There are approximately 20,000 to 50,000 genes in the human genome, according to various estimates. In humans, genes vary in size from few hundred bases to 200 kilobases (1 kb = 1000 bases) but some of them span over more than 2000 kb. Most but not all of the DNA sequence is identical in all humans. The sequence of nucleic acids composing a gene is called allele. A locus designs a specific location on the genome, either at a single base or a sequence of bases. As genes are associated with a precise location on the genome, both terms are often mixed up, even though their precise definition is different. A gene is said to be polymorphic if two or more alleles exist for it in proportions greater than 1%. The wild-type allele is the most commonly found in general population, although many genes are polymorphic with several commonly found alleles. An individual is said to be homozygous (respectively heterozygous) for a given gene if both his paternal and maternal alleles are identical (respectively different).

A gene is composed of alternate sequences of exons and introns. A regulatory region working as a switch on and off button is located upstream each gene in the 5' direction and contains the promotor of the gene. An abnormal regulation of promotors of certain genes is often linked to cancer.

Most genes are expressed in three phases: transcription, RNA maturation and translation. During the transcription of a gene, its DNA sequence is used as a pattern to build its complementary sequence of ribonucleic acid (RNA). Bases in RNA are the same as bases in DNA except thymidine which is replaced by uracile (U). Transcription is initialized when the promotor binds a transcription factor. A given transcription factor is specific to a DNA sequence hence it is able to active only genes containing that sequence in its regulatory site. Transcription termination involves different processes including multiple adenosine added to the 3' end of the RNA. Maturation of RNA consists in various modifications including alternate splicing of introns in order to prepare RNA for translation. The resulting molecule is called messenger RNA (mRNA). Translation is a process during which a sequence of mRNA is used as a template to build a sequence of amino-acids called a protein according to the *genetic code* given in Table 3.1. This table is an extraction from the NCBI (National Center for Biotechnology Information) website[1]. The genetic code is decoded by trios of nucleotides called codons, each one coding either for an amino-acid or a stop signal. As 64 codons code for 22 amino-acids and a stop, the genetic code is redondant and is said to be degenerated. Hence, as detailed in a deeper sense in Section 3.1.2.4, some mutations have no impact on the resulting amino-acid sequence and are said to be *silent*.

3.1.2.2 Cell cycle and mitosis

Most cells alternate cell cycles divided in four ordered phases: Gap 1 (G1), Synthesis (S), Gap 2 (G2) and Mitosis (M). An extra phase called Gap 0 or quiescence may be entered after G1 by non-dividing cells. Quiescence is common for fully differentiated cells (such as most neurons) and lasts long periods of time, possibly indefinitely.

[1] `https://www.ncbi.nlm.nih.gov`

SECOND POSITION

FIRST POSITION	U	C	A	G	THIRD POSITION
U	phenyl-alanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
	* methionine		lysine	arginine	A
					G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

Figure 3.1: The genetic code (source: National Center for Biotechnology Information).

Each passage from a phase to the next is controlled by checkpoints which ensure that the cell passed properly a phase and is ready for the next one. Abnormalities in genes coding for proteins involved in checkpoints are often linked to cancer.

The longest phase is G1 during which chromosomes are diffuse in order to allow for translation and transcription into proteins. Whenever a division is required and checkpoints passed, the cell enters into the S phase during which DNA is duplicated. The double helix is opened and each strand is used as a pattern to synthesize its complementary. Each resulting chromosome hence consists in two identical chromatids linked together in a region called centromere. Errors during S phase may lead to mutations. The cell enters then briefly the growth phase G2. If the cell passes the checkpoints between G2 and M (mainly controlled by the protein p53), it enters into mitosis represented in Figure 3.2. This figure is extracted from the NHGRI (National Human Genome Research Institute) website[2]. Mitosis starts with the compaction of chromosomes into an H structure which is the only visible form under microscope, hence the way scientists represent chromosomes in karyotypes. Each chromosome composed of two identical chromatids is split in two and each sister chromatid migrates to an opposite pole of the cell. The cell is divided in two daughter cells containing each original chromosome composed of one chromatid. Hence the two diploid daughter cells contain (normally) identical genome.

[2] `https://www.genome.gov`

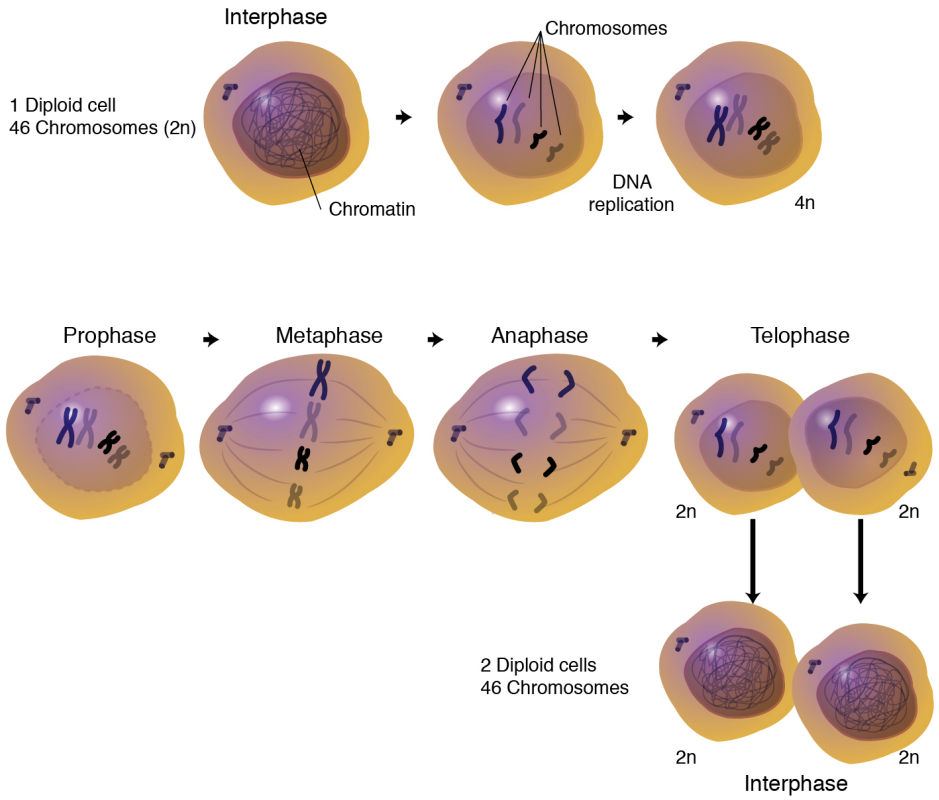


Figure 3.2: Graphical representation of a mitosis (source: National Human Genome Research Institute).

3.1.2.3 Meiosis

A meiosis is another type of cell division occurring solely in germ cells and represented in Figure 3.3 extracted from the NHGRI website. It is composed of two division phases, meiosis I and meiosis II, resulting, from one diploid cell, to four haploid gametes (sperm or egg cells). A meiosis is preceded by a growth phase and an S phase leading to chromosomes composed of two sister chromatids attached by a centromere. During meiosis I, homologous chromosomes are separated. Each chromosome of a paire migrates to an opposite pole of the cell and the cell split in two. The assigned pole is random leading to a minimum of 2^{23} possible combinations. In fact this number is greatly increased by a phenomenon called crossing-over detailed in the next paragraph. Each daughter cell progresses into meiosis II during which centromeres are split and each sister chromatid migrates to an opposite pole of the cell. Cells are split in two resulting each in two haploid cells. Hence one diploid cell entering in meiosis gives four haploid cells called gametes.

A crossing-over is a phenomenon occurring during meiosis I. Two homologous chromosomes pair up and exchange some genetic material leading to recombinant gametes at the end of the meiosis as represented in Figure 3.4 extracted from the NHGRI website. As a result, crossing-overs allow for an increase of genetic diversity. Crossing-overs tend to appear more frequently in chromosomal regions called hot spots. The closer two genes are located on the same chromosome, the smaller the probability of their alleles to be separated during gametogenesis. Two genes are said to be in linkage disequilibrium if the combinations of their alleles (called *haplotypes*) are different from the one that would be expected if they were randomly distributed. This postulate opened doors to the definition of genetic distance between genes first introduced by Thomas Morgan and his colleagues and further developed by John Burdon Sanderson Haldane in his paper Haldane (1919) (see Section 3.1.3.2 for an introduction to genetic linkage studies which aim at localizing genes on the genome).

3.1.2.4 Mutation

A mutation is a brutal modification of the genome most of the time caused by errors during DNA replication, mitosis or meiosis. It is a stochastic phenomenon whose frequency is increased by factors called mutagens (pollution, X-ray, inappropriate diet, etc.). Large-scale mutations (respectively small-scale mutations) also called chromosomal rearrangement affect a large portion of chromosomes (respectively one or few nucleotides). A small-scale mutation affecting one single nucleotide is called a point mutation.

Large-scale mutations include duplications (addition of an extra segment of chromosome to another chromosome), deletions of a segment of chromosome, inversions (180 degree reverse orientation of a segment of chromosome), translocations (exchange of segments between non-homologous chromosomes). Small-scale mutations include substitutions of one nucleotide by another, insertions and deletions of one or few nucleotides.

Small-scale mutations are classified as neutral, deleterious, beneficial or nearly

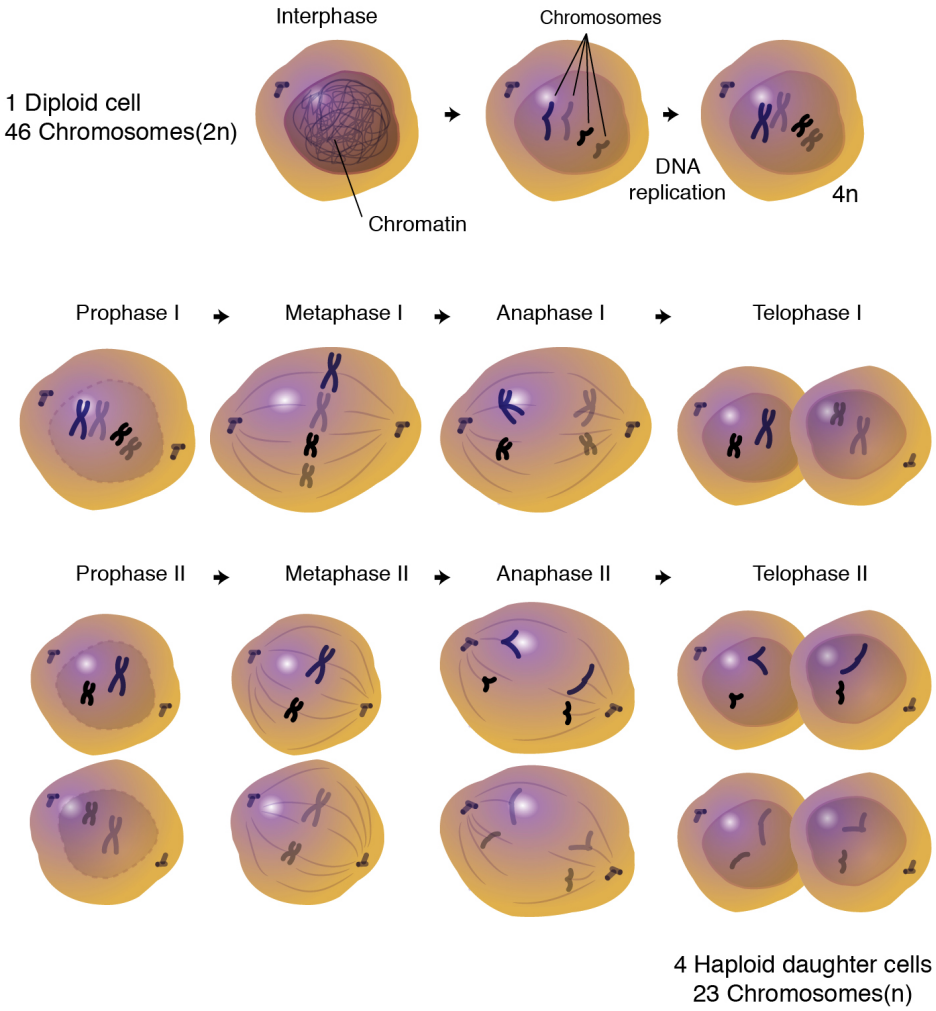


Figure 3.3: Graphical representation of a meiosis (source: National Human Genome Research Institute).

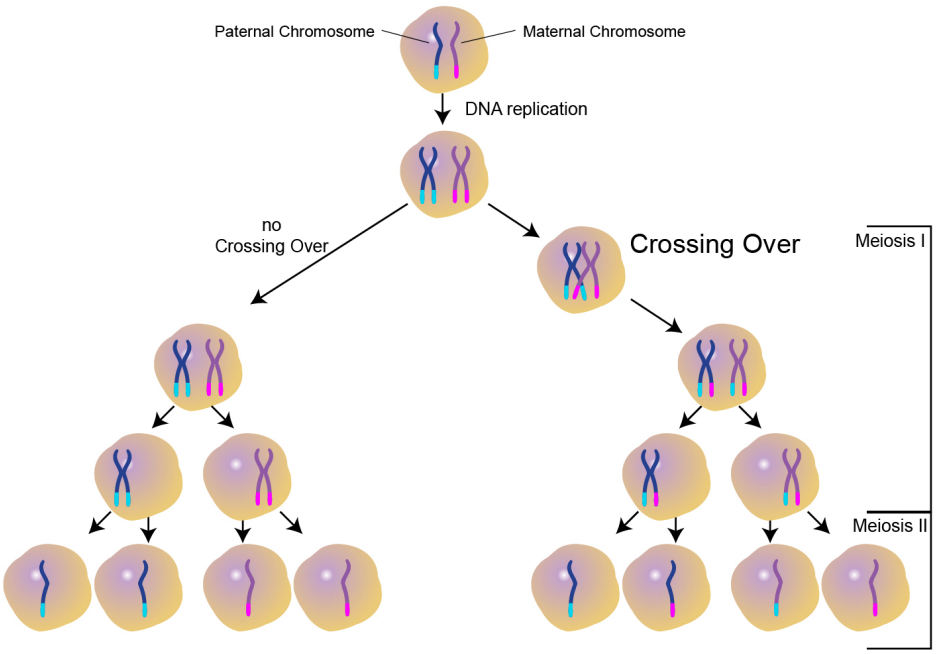


Figure 3.4: Graphical representation of a crossing-over (source: National Human Genome Research Institute).

neutral according to their impact on the protein production, composition and function, and the health of the organism. Nearly neutral mutations are ignored in most genetic models. An insertion or a deletion is said to be frameshift (respectively in frame) if it concerns a number of nucleotides not divisible by three (respectively divisible by three). Because of the triplet nature of the genetic code, a frameshift mutation results in a completely different sequence of amino-acids. An inframe mutation may vary from neutral to deleterious (or beneficial). A substitution is said to be silent (respectively nonsens and missens) if the resulting codon codes for the same amino-acid as a result of the degenerescence of the genetic code (see Table 3.1) (respectively a stop signal and another amino acid).

A mutation affecting gametal DNA can be transmitted to an offspring who carries it constitutionally (in all his cells). It can therefore be transmitted again to the next generation. A mutation acquired constitutionally is said to be *inherited*. A mutation is said to be *somatic* if it is not inherited but acquired in life and only found in daughter cells of the cell firstly affected. A *de novo* mutation is a mutation carried constitutionally by an individual but not by his/her parents and is usually a consequence of a somatic mutation in parental gametes and less frequently a *post-zygotic* mutation which appears in the egg after fertilization and is expressed in a mosaic form.

3.1.3 Patterns of heredity and gene expression

3.1.3.1 Dominance, penetrance, mode of inheritance

The genotype of an individual is defined as the set of paternal and maternal alleles received from his parents and can be restricted to one or few genes. We define a major (respectively minor) gene to be a gene whose effect has a major (respectively minor) impact on a phenotype and a polygenic factor to be the combination of multiple genetic components, each having a small effect, but the total effect is statistically significant. Deleterious mutations in major genes are rare and for most of theme a double homozygous mutated genotype is either extremely rare or lethal (before birth). A familial aggregation is defined to be a systematic tendency for a trait to cluster in families. It may have one or multiple causes among which one or several major gene(s), a polygenic effect, shared environmental factors, etc., acting independently or in interactions. Complex diseases result of complex interactions between multiple genes and/or genes and environmental factors and involve an age-dependent expression.

Mendelian inheritance. Before the discovery of genetic material, Gregor Mendel was the first to propose and gave his name to a coherent pattern of inheritance called *Mendelian inheritance*, *Mendel's laws of inheritance* or *Mendel's principles of inheritance*. Conducting experiments on peas, observing seven different traits and assuming that each trait is discrete, Mendel proposed the following three laws:

- Law of dominance and uniformity: each trait is expressed in different forms (called alleles today) and an individual carries two forms, each coming from one parent. Some traits have a dominant form over others. The dominant form masks the other one and is expressed in the trait.
- Law of segregation: a gamete carries one form for each trait. The segregation of the two forms during gamete formation is random and uniform.
- Law of independent assortment: Traits are transmitted independently.

Mendel recognized himself that the postulates he made only apply to certain types of species and traits and they fail in characterizing complex traits in most species. However he proposed the first bricks of gene inheritance, later refined and developed, but still constitutive of the foundations of modern laws of inheritance.

Penetrance and dominance. Let d and D be two different alleles of a gene involved in the expression of a phenotype Y, we say that

- D is dominant over d if $\mathbb{P}(Y|dd) \neq \mathbb{P}(Y|dD)$ and $\mathbb{P}(Y|dD) = \mathbb{P}(Y|DD)$ (if DD is not lethal). In other words a single copy of D is sufficient for its causal relation towards the phenotype Y.
- D is recessive over d if $\mathbb{P}(Y|dD) = \mathbb{P}(Y|dd)$

- d and D are codominant if $\mathbb{P}(Y|dd) \neq \mathbb{P}(Y|dD) \neq \mathbb{P}(Y|DD)$
- d and D have an additive (respectively multiplicative) effect if the contribution of each allele in heterozygous genotypes is added (respectively multiplied).

We define the penetrance to be the probability density function of the phenotype conditional on the genotype.

Mode of inheritance. The mode of inheritance refers to the location of a major gene and dominance of his alleles. Assuming a dichotomous classification of the different alleles as non-deleterious (d) or deleterious (D), we distinguish the following modes of inheritance: autosomal dominant, autosomal recessive, X-linked dominant, X-linked recessive, Y-linked and mitochondrial. A mitochondrion is an organelle containing DNA found in most eukaryotic cells, essentially involved in energy production and transmitted from mother to offspring. Mitochondrial inheritance applies to genes located on mitochondrial genome. In this thesis, we restrict our work on autosomal dominant and autosomal recessive mode of inheritance applying, as their names speak for themselves, to genes located on autosomes whose alleles respectively act in a dominant and recessive manner. Most genetic models assume independence between parental origin and expression of an allele, ie. heterozygous genotypes dD have same penetrance regardless the paternal or maternal origin of d and D. However rare exceptions called *genomic imprinting* have been pointed out in the last decades. Under Mendelian assumption of independent segregation of alleles during gametogenesis and no genomic imprinting, the mode of inheritance fully defines the probability density function of genotypes of an offspring conditional on the genotypes of his parents in case of a single gene or several independent genes. However, when considering several genes in linkage disequilibrium, the Mendelian pattern does not apply and the segregation of alleles of different genes are functions of the genetic distance between genes.

Hardy Weinberg equilibrium. The Hardy-Weinberg (HW) equilibrium is a theorem in population genetics which states that allele and genotype frequencies in a population are invariant under the following conditions: infinite size of the population, panmixia (random union of individuals), pangamia (random union of gametes), no migration, no mutation, no natural selection and discrete successive generation. HW equilibrium is easily proved in the simple case of a biallelic locus, with alleles d and D and respective frequencies $1-q$ and q at a generation. Under listed above conditions, the frequency of allele d in the next generation is given by $f(d) = f(dd) + \frac{1}{2}f(dD) = (1-q)^2 + q(1-q) = 1-q$. Similarly $f(D) = f(DD) + \frac{1}{2}f(dD) = q$. Deviations from HW conditions are commonly ignored in genetic analyses and most genetic models assume HW equilibrium for founders (individuals with no ancestor).

3.1.3.2 Introduction to genetic epidemiology

The study of genetic components in genetic and complex diseases, their interactions and interplay with environmental factors is a field of statistical genetics called ge-

netic epidemiology (see Thomas et al., 2004, for an introduction to main statistical methods in genetic epidemiology). A Mendelian disease is a disease caused by deleterious mutations in a single gene. It is now well recognized that complex diseases have multiple possible causes. Causing factors of breast cancer for instance include two major multiallelic genes, several minor genes, polygenic factors, environmental factors and Gene × Gene and Gene × Environment interactions are involved. However a theoretical framework is needed and the concept of causal factors is studied as an age-dependent increase of risk of developing a disease, all other factors remaining constant. Whereas the concept of causation is essential in genetics, counterfactual factors and simply associations must be distinguished as discussed in (Page et al., 2003).

Genetic epidemiology is divided in several steps as detailed in (Thomas et al., 2004) and summarized below.

- **Familial aggregation** questions whether there exists an evidence of aggregation of disease within families and if the pattern is consistent with a genetic influence. Note that a familial aggregation does not necessarily imply a genetic factor as families share more than genes. Therefore familial aggregation studies combine twin and adoption studies in order to estimate the degree of environmental and genetic effect on a trait. Twin studies compare monozygotic twins who where originally the same zygote split in two and share identical genome and dizygotic twins who shared the same womb but are developed from two different ovum fertilized by two different sperm cells. Monozygotic and dizygotic twins are assumed to share the same environmental factors. Adoption studies compare parents and their biological offsprings they raised, biological parents and their offsprings raised appart, adoptive parents and their adopted offsprings they raised.

- **Segregation analysis** comes next and aims at determining the genetic pattern (one or more major gene and/or polygenic factors) and mode of inheritance as well as estimating parameters introduced in Section 3.1.3.1. Pedigree-based genetic models are fit to data on phenotypes of family members (see Section 3.2 for an introduction to pedigree-based models) and the most likely mode of inheritance along with parameters are estimated using maximum likelihood estimation.

- **Linkage analysis** aims at localizing a major gene, previously pointed at by segregation analyses, on the genome, using the phenomenon of crossing-overs during gametogenesis. A Single Nucleotide Polymorphism (SNP pronounced snip) is the substitution of a single nucleotide by another at a specific and known location on the genome and in proportions greater than one per cent in the general population. SNPs are the most commonly known polymorphism. As their location is known on the genome, they are used as markers in order to localize genes of interest by estimating the proportion of recombinant gametes in pedigrees using pedigree-based models. Two-point linkage (respectively multi-point linkage) aims at estimating the proportion of recombinant gametes in a set of pedigrees between one SNP (respectively multiple SNPs)

and a gene of interest. In two-point linkage, the parameter θ is defined as the probability of an odd number of crossing-overs between the marker and the gene of interest, ie. the proportion of recombinant gametes (see Lauritzen and Sheehan, 2003, section 4.1). The null hypothesis $\theta = 0.5$ is tested against $\theta < 0.5$ using usually a function of a likelihood ratio test called the LOD score of θ defined as $\text{LOD}(\theta) = \log_{10}(L(\theta)/L(0.5))$, where $L(\theta)$, the likelihood of θ, is evaluated at $\hat{\theta}$, an estimate of θ using maximum likelihood estimation. Assuming that the distribution of the number C of crossing-overs between the two loci is a Poisson distribution, the genetic distance δ between the two loci called is defined as its parameter, hence, the expected number of crossing-overs . Its unitary measure is given in Morgan (and more often in centiMorgan) in honor of Thomas Morgan and his colleagues who discovered the non-independent segregation of genes. The Haldane's mapping function detailed below links θ and δ:

$$\theta = \sum_{n=0}^{+\infty} \mathbb{P}(C = 2n+1) = \sum_{n=0}^{+\infty} \frac{e^{-\delta}\delta^{2n+1}}{(2n+1)!} = e^{-\delta}\left(\frac{e^{\delta} - e^{-\delta}}{2}\right) = \frac{1 - e^{-2\delta}}{2} \quad (3.1)$$

leading to $\delta = -\dfrac{\log(1-2\theta)}{2}$. Multipoint linkage analysis (Ott, 1999) involve multiple markers and increase computational complexity as we will see in Section 3.2.

- **Association studies** are population based, hence they do not require pedigrees nor the specification of a disease model. They study the association between candidate genes and a phenotype or between the whole genome (Genome Wide Association Studies - GWAS) and a phenotype. They include analyses of Gene $\times$ Gene and Gene $\times$ Environment interactions. The rapid and recent development of sequencing methods provided an extensive interest in association studies in recent years. As this thesis focuses on pedigree-based models, association studies are outside its scope.

3.2 Pedigree-based models

Pedigree-based models are Bayesian Networks (BNs) extensively used in complex genetic problems and in particular for estimating parameters in segregation and linkage analysis (see Section 3.1.3) and for computing risks of genetic predisposition or disease risks in genetic counseling. In this section we propose an introduction to these models. We start with the definition of a pedigree (Section 3.2.1), we pursue with an overview of main tools for its implementation in a BN (Section 3.2.2) and we finish with a list of principal inferences performed over these graphs and a review about main existing algorithms in genetic analyses (Section 3.2.3). In this whole section, L, N and n respectively stand for the number of studied diseases (L), the number of genes involved (N) and the number of individuals, that is family members, in the pedigree (n).

3.2.1 Definition of a pedigree

A pedigree is a graphical representation of a set of family members indicating their sex, parental relationships and phenotypes according to the study. It is drawn using standard formatting and nomenclature as pictured in Figure 3.6 extracted from the National Cancer Institute website[3]. Each generation is usually represented on a same row. The individual (usually affected) who seeks medical attention by a genetic counselor and who is the reason why the family is studied is called the *proband* and is denoted by an arrow pointing at him. An individual with no reported ancestor is called a *founder* and we denote by $\mathcal{F}$ (respectively $\overline{\mathcal{F}}$) the set of indexes for founders (respectively for non-founders).

As seen in Section 3.1.1, a phenotype can be of various form (dichotomous, categorical, continuous, a time-to-event data, etc.). Chapter 8 will be devoted to the development of a pedigree-based model in the framework a genetic predisposition to cancer called the Lynch syndrome. Phenotypes in that model are time-to-event and sets of biological and clinical data. In this introductory section, for the sake of simplicity, we restrict phenotypes to time-to-event data. A parametric survival model such as one of those introduced in Sections 2.2 or 2.3 is chosen according to the context. Let us consider, in this introductory section, to the competing risk model as the one represented in Figure 2.2 and recalled in Figure 3.5 with the notation used in the present section, for it to be implemented in the only currently exiting pedigree-based model for the Lynch syndrome called MMRpro (Chen et al., 2006). We will see in Chapter 8 more complexe multi-state models involving multiple transitent states for modeling the evolution of a patient through different cancer types.

We denote by $\mathcal{Z} = \{\text{UN}, D^1, \ldots, D^L\}$ the state space such that UN stands for "Unaffected" and for $d \in \{1, \ldots, L\}$, D^d stands for "Diagnosed with disease D^d". Let Z_i be the phenotype of individual $i \in \{1, \ldots, n\}$ in a family on n members, Z_i is a stochastic process $\{Z_i(t), t \geq 0\}$ where $Z_i(t)$ denotes the state occupied by individual

[3] `https://www.cancer.gov/publications/dictionaries/genetics-dictionary/def/pedigree`

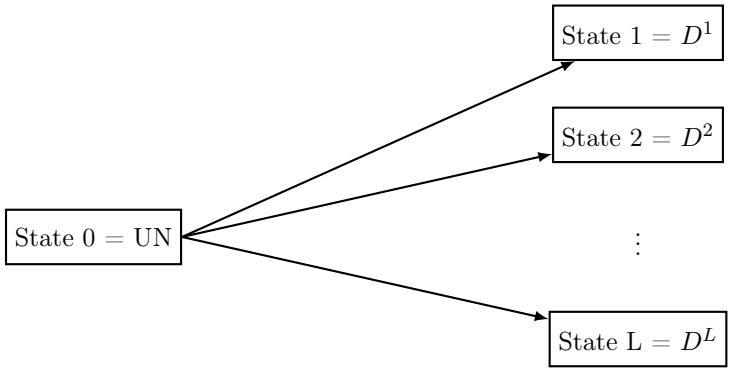


Figure 3.5: Competing risk model.

i at time t. Each (observed) phenotype is written below the associated individual on the pedigree drawing using simplified notation. Let PH_i (personal history of individual i) be the phenotype of individual i in its simplified version, we have for instance $\{\mathrm{PH}_i = \mathrm{UN}_t\} \equiv \{Z_i(t) = \mathrm{UN}, t \geq 0\}$ for an individual free of disease from birth up to age t; for $d \in \{1, \ldots, L\}$, $\{\mathrm{PH}_i = D^d_t\} \equiv \{Z_i(s) = \mathrm{UN}, s < t, Z_i(t) = D^d\}$ for an individual free of disease up to age t and diagnosed with D^d at age t; for $d, e \in \{1, \ldots, L\}$, s < t, $\{\mathrm{PH}_i = \{D^d_s, D^e_t\}\} \equiv \{Z_i(u) = \mathrm{UN}, u < s, Z_i(v) = D^d, v \in [s, t[, Z_i(t) = D^e\}$ for an individual free of disease up to age s, diagnosed with D^d at age s and D^e at age t, free of any other studied disease; $\{\mathrm{PH}_i = \{D^d_r, D^e_s, \mathrm{DCD}_t\}\} \equiv \{Z_i(u) = \mathrm{UN}, u < r, Z_i(v) = D^d, v \in [r, s[, Z_i(w) = D^e, w \in [s, t]\}$ for an individual free of disease up to age r, diagnosed with D^d at age r and D^e at age s, deceased at age t, free of any other studied disease. As the simplified notation is the one commonly used in genetic counseling, for the rest of the thesis, we will often use it, especially in applied sections. A phenotype may be partially reported such as an uncertain disease type or a time interval instead of age at diagnosis or censoring.

An example of a pedigree with three studied diseases, colon, rectum and endometrium cancer, respectively denoted CC, RC and EC is drawn in Figure 3.7. Deviations to standard formatting may be inevitable in some complex mating as represented Figure 3.8 with two equivalent pedigrees, one including the duplication of an individual and the other one representing non-aligned individuals of same generation.

3.2.2 Bayesian networks in genetic analyses

As explained in Lauritzen and Sheehan (2003) and Koller and Friedman (2009), probabilistic graphical models are particularly appropriate and extensively used in pedigree analyses for modeling probabilistic relationships between variables. We detail in this section the implementation of a pedigree into a Bayesian network (BN) assuming that phenotypes are coded by autosomes. Sex-linked inheritance will not be considered in this thesis but involves similar reasoning with appropriate dependency structure and Conditional Probability Distributions (CPDs).

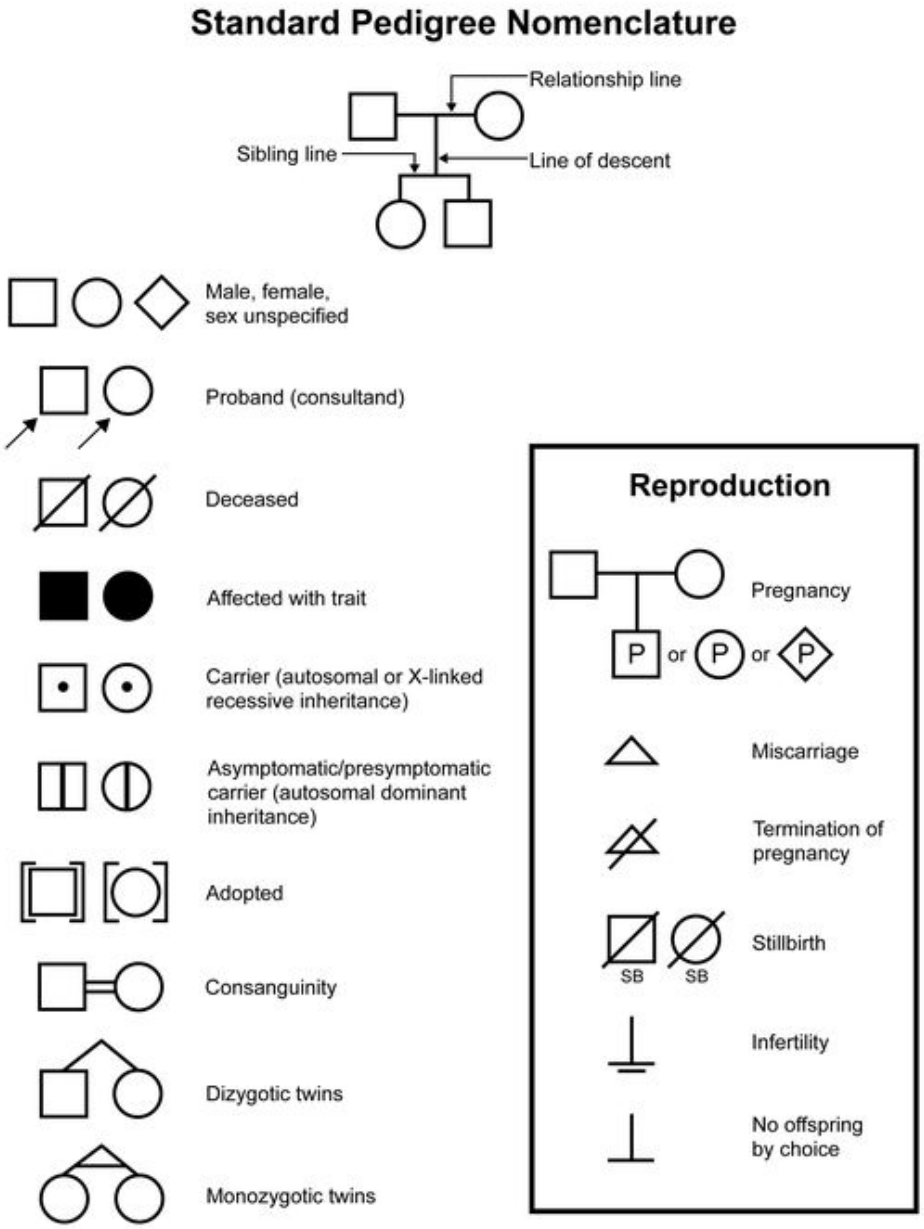


Figure 3.6: Standard pedigree nomenclature (source: National Cancer Institute).

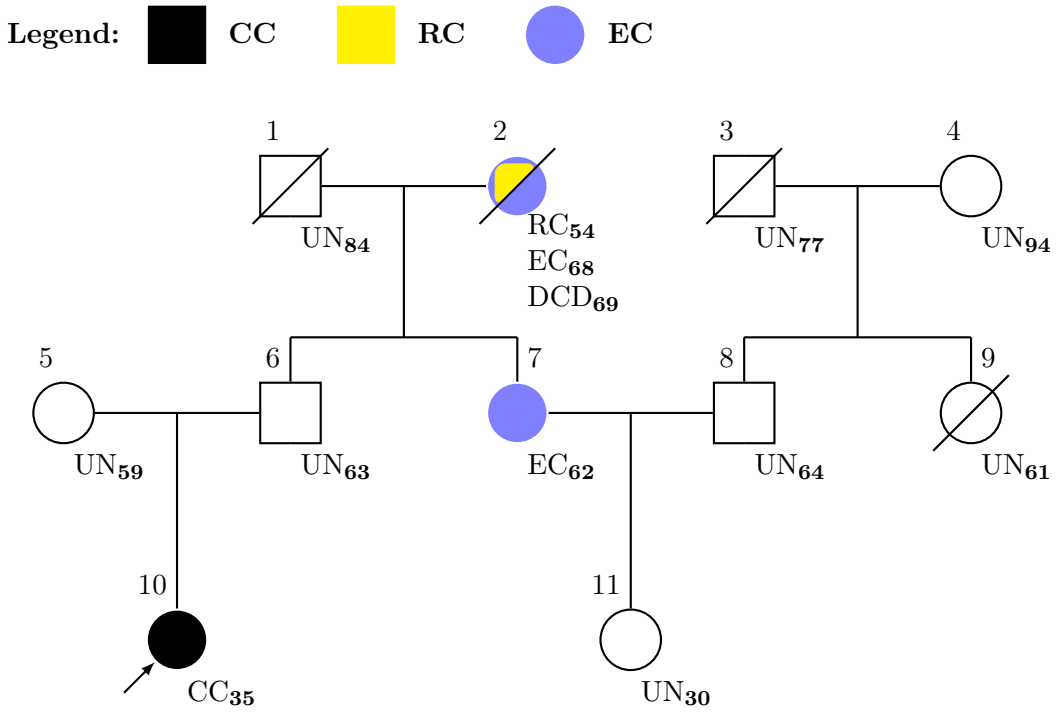


Figure 3.7: An example of a pedigree with time-to-event data as phenotypes and three studied diseases: colon, rectum and endometrium cancer respectively denoted CC, RC and EC. Disease types are denoted by a specific color.

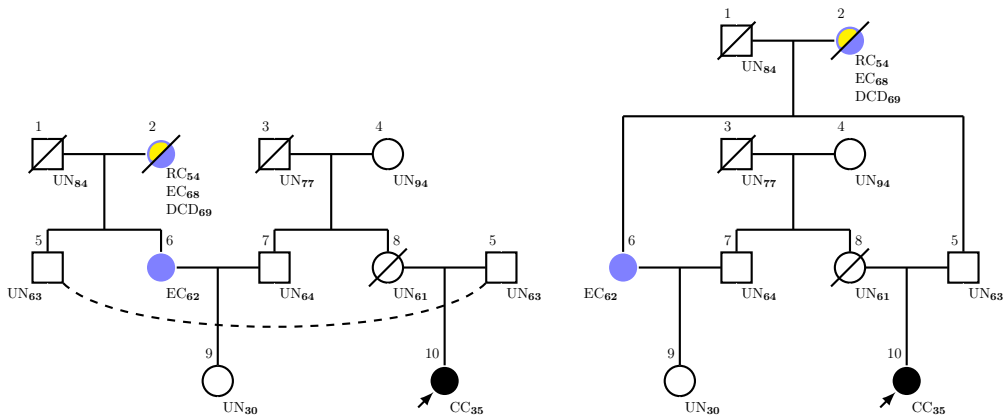
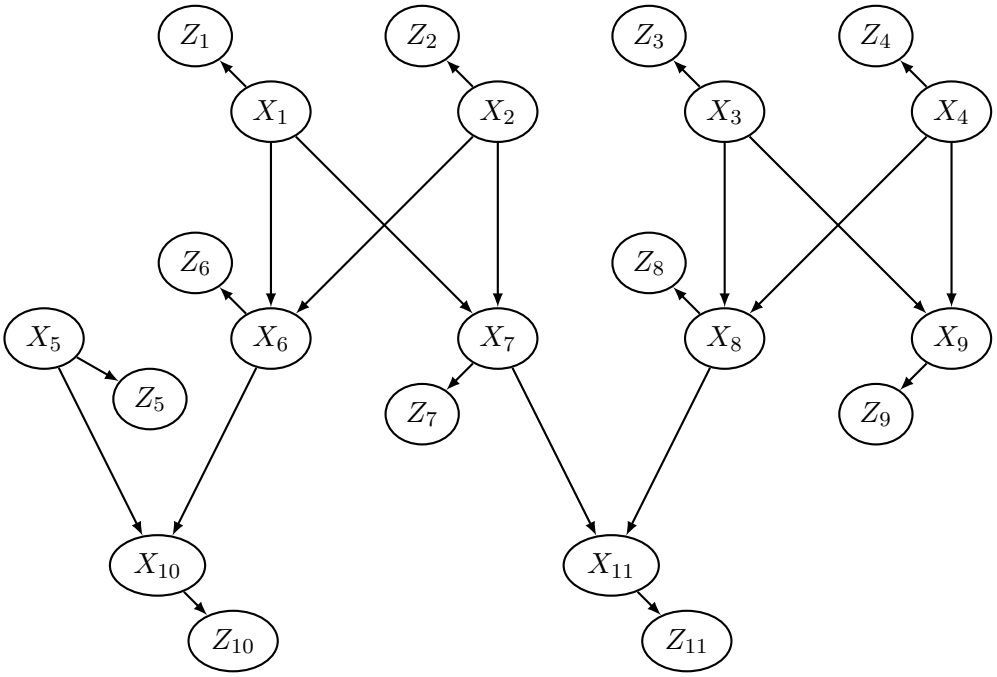
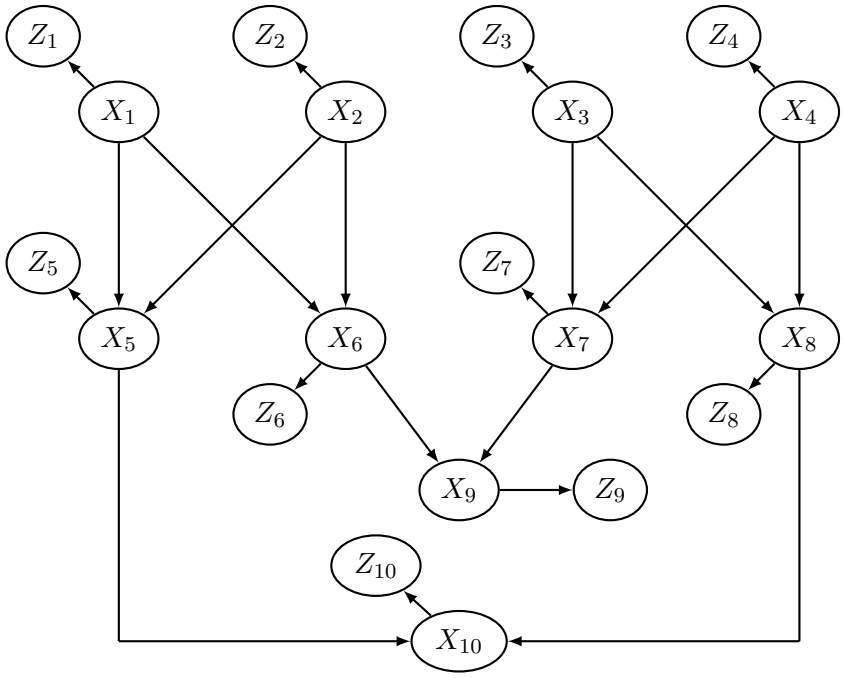


Figure 3.8: An example of a pedigree with complex mating whose representation involves either the duplication of Individual 5 (on the left) or non-aligned Individuals 1, 2, 3 and 4 of same generation (on the right). Disease types are denoted by a specific color.



(a) Genotype DAG associated with the pedigree represented in Figure 3.7



(b) Genotype DAG associated with the pedigree represented in Figure 3.8

Figure 3.9: Examples of genotype DAGs where X_i denotes the genotype carried by individual $i \in \{1, \ldots, n\}$ and Z_i, his/her phenotype.

3.2.2.1 Genotype Bayesian Network

The most intuitive BN associated with a pedigree is a genotype BN defined as $\mathcal{B}^{\text{geno}} = (G^{\text{geno}} = (\{X, Z\}, \mathcal{E}^{\text{geno}}), \mathbb{P})$ where $X = \{X_i\}_{i\in\{1,\dots,n\}}$ is the set of genotypes, $Z = \{Z_i\}_{i\in\{1,\dots,n\}}$ is the set of phenotypes, $\mathcal{E}^{\text{geno}} = \{(X_{\text{p}(i)}, X_i)_{i\in\overline{\mathcal{F}}}, (X_{\text{m}(i)}, X_i)_{i\in\overline{\mathcal{F}}}, (X_i, Z_i)_{i\in\{1,\dots,n\}}\}$ where, for all $i \in \overline{\mathcal{F}}$, $\text{p}(i)$ (respectively $\text{m}(i)$) is the index of the father (respectively mother) of individual i and $\mathbb{P}$ is a probability distribution detailed below. A graphical representation of two genotype DAGs associated with pedigrees drawn in Figure 3.7 and Figure 3.8 is proposed in Figure 3.9. The Markov dependence between X_i and Z_i for all $i \in \{1, \dots, n\}$ is straightforward. However the Markov dependence, for $i \in \overline{\mathcal{F}}$, between X_i and $X_{\text{p}(i)}$ as well as X_i and $X_{\text{m}(i)}$ is verified if alleles segregate independently. The joint probability of X and Z is given by the following product of CPDs:

$$\mathbb{P}(X, Z|S, \theta) = \prod_{i=1}^{n} \underbrace{\mathbb{P}(X_i|X_{\text{p}(i)}, X_{\text{m}(i)}; \theta^x)}_{\substack{\text{inheritance / genotypic} \\ \text{component}}} \underbrace{\mathbb{P}(Z_i|X_i, S_i; \theta^z)}_{\substack{\text{phenotypic} \\ \text{component}}} \tag{3.2}$$

where, for all $i \in \mathcal{F}$, $\text{p}(i) = \text{m}(i) = 0$, $X_0 = \emptyset$ by convention and $S = \{S_i\}_{i=1,\dots,n}$ where S_i is a binary variable denoting the sex of individual i and takes value 1 for males and 2 for females. The parameter θ^x includes allele frequencies in general population, mode of inheritance, etc., and θ^z includes dominance, penetrance, genomic imprinting, etc. (see Section 3.1.3).

Genotypic component Most pedigree-based models used in medical genetics and genetic counseling are built under the following assumptions.

A 1. *Biallelic genes such that their component alleles take value* 0 *if non-pathogenic and* 1 *otherwise*

A 2. *Hardy-Weinberg equilibrium for genotypes of founders*

A 3. *Independent segregation of alleles per gene*

A 4. *No genomic imprinting*

Furthermore we assume in this section that genes segregate independently, i.e. they are not in linkage disequilibrium. Linkage disequilibrium will be considered in the section devoted to selector BNs (Section 3.2.2.3). Let $\mathcal{G} = \{G^1, \dots, G^N\}$ be a set of N genes segregating independently, under Assumptions A1, A2 and A3, the parameter θ^x is reduced to $q = (q_g)_{g=1,\dots,N}$, the vector of pathogenic allele frequencies such that, for all $g \in \{1, \dots, N\}$, q_g is the frequency of pathogenic alleles for gene G^g in the general population. If no genotype is lethal (i.e. leads to death before birth), under Assumption A1, the state space of X_i is $\mathcal{X} = \{00, 01, 10, 11\}^N$. It would be, for instance, reduced to $\{00, 01, 10\} \times \{00, 01, 10, 11\}^{N-1}$ if the genotype homozygous carrier for G^1 where lethal. We assume in this introductory section that no genotype is lethal. Under Assumption A4, the origin (paternal or maternal) of an allele has no influence on its effect on the phenotype, hence, $\mathcal{X}$ can be reduced to

	$X_{\mathrm{p}(i)} = 00,\ X_{\mathrm{m}(i)} = 00$	$X_{\mathrm{p}(i)} = 01,\ X_{\mathrm{m}(i)} = 00$	$X_{\mathrm{p}(i)} = 11,\ X_{\mathrm{m}(i)} = 00$
$X_i = 00$	1.0	0.5	0.0
$X_i = 01$	0.0	0.5	1.0
$X_i = 11$	0.0	0.0	0.0
	$X_{\mathrm{p}(i)} = 00,\ X_{\mathrm{m}(i)} = 01$	$X_{\mathrm{p}(i)} = 01,\ X_{\mathrm{m}(i)} = 10$	$X_{\mathrm{p}(i)} = 11,\ X_{\mathrm{m}(i)} = 01$
$X_i = 00$	0.5	0.25	0.0
$X_i = 01$	0.5	0.5	0.5
$X_i = 11$	0.0	0.25	0.5
	$X_{\mathrm{p}(i)} = 00,\ X_{\mathrm{m}(i)} = 11$	$X_{\mathrm{p}(i)} = 01,\ X_{\mathrm{m}(i)} = 11$	$X_{\mathrm{p}(i)} = 11,\ X_{\mathrm{m}(i)} = 11$
$X_i = 00$	0.0	0.0	0.0
$X_i = 01$	1.0	0.5	0.0
$X_i = 11$	0.0	0.5	1.0

Table 3.1: CPDs $\mathbb{P}(X_i|X_{\mathrm{p}(i)}, X_{\mathrm{m}(i)})$ for non-founders in the framework of a single biallelic gene X with no lethal genotype.

$\{00, 01, 11\}^N$. Let X_i^g be the genotype of individual i for gene G^g, under Assumption A2, for all $i \in \mathcal{F}$ (set of founders), for all $g \in \{1, \ldots, N\}$, $\mathbb{P}(X_i^g = 00|q_g) = (1 - q_g)^2$, $\mathbb{P}(X_i^g = 01|q_g) = 2q_g(1 - q_g)$, $\mathbb{P}(X_i^g = 11|q_g) = q_g^2$. We have $\mathbb{P}(X_i|q) = \prod_{g=1}^N \mathbb{P}(X_i^g|q_g)$, an CPD of cardinality 3^N under above assumptions. Moreover, for all $i \in \overline{\mathcal{F}}$, CPDs of the form $\mathbb{P}(X_i|X_{\mathrm{p}(i)}, X_{\mathrm{m}(i)})$ are of cardinality $|\mathcal{X}|^3$, i.e. 3^{3N} under above assumptions. Hence, a genotype BN is intuitive but leads to high computational complexity especially with an increasing number of genes N. An allele BN detailed in the Section 3.2.2.2 is usually preferred. Values taken by CPDs of the form $\mathbb{P}(X_i|X_{\mathrm{p}(i)}, X_{\mathrm{m}(i)})$ are reported in Table 3.1 under Assumption A3 and in the framework of a single gene (N=1) for the sake of readability.

Phenotypic component CPDs of the form $\mathbb{P}(Z_i|X_i, S_i; \theta^z)$ are parametrized by θ^z and computed according to the chosen multi-state model. As previously mentioned, we consider in this section the competing risk model drawn in Figure 3.5. A variety of extensions exist including polygenic effects (Antoniou et al., 2002), frailty model (Gorfine et al., 2013), etc. Covariates can also be added. Considering L diseases $D^1, \ldots, D^L$, State 0 stands for "Unaffected" and for $k \in \{1, \ldots, L\}$, state k stands for "Diagnosed with disease D^k". In such models, diseased states are all absorbant, hence multiple diagnoses can not be taken into account and the first diagnosis is the only one considered. Transition intensities are sex and genotype dependent and parametrized by θ^z. We denote by $\lambda_k^{s,x}$, the transition intensity from state 0 to state $k \in \{1, \ldots, L\}$ conditional on sex $s \in \{1, 2\}$ and genotype $x \in \mathcal{X}$. Let T_i^k be the time at first diagnosis of disease D^k and T_i^* be the time at first diagnosis of any disease in the set $\{D^1, \ldots, D^L\}$, we detail thereafter the expression of CPDs of the phenotypic component in a competing risk model represented in Figure 3.5

using the simplified notation introduced in Section 3.2.1:

$$\begin{aligned}\mathbb{P}(\mathrm{PH}_i = \mathrm{UN}_t | X_i = x, S_i = s; \theta^z) &= \mathbb{P}(T_i^* > t | X_i = x, S_i = s; \theta^z) \\ &= \exp\left(- \int_0^t \sum_{k=1}^{L} \lambda_k^{s,x}(u) \mathrm{d}u \right) \end{aligned} \tag{3.3}$$

and

$$\begin{aligned}\mathbb{P}(\mathrm{PH}_i = \mathrm{D}_t^k | X_i = x, S_i = s; \theta^z) &= \mathbb{P}(T_i^k = t | X_i = x, s_i = s; \theta^z) \\ &= \exp\left(- \int_0^t \sum_{\ell=1}^{L} \lambda_\ell^{s,x}(u) \mathrm{d}u \right) \lambda_k^{s,x}(t) \end{aligned} \tag{3.4}$$

where, for $k \in \{1, \ldots, L\}$, $\mathbb{P}(T_i^k = t | X_i = x, S_i = s; \theta^z)$, called the penetrance of genotype x for disease D^k for individuals of sex $s \in \{1, 2\}$, is the density of T_i^k conditional on $S_i = s$ and $X_i = x$.

3.2.2.2 Allele Bayesian network

In an allele BN, genotypes are split into their component alleles in order to reduce computational complexity. Indeed as detailed in Chapter 1, exact inferences with the sum-product algorithm are linear in the number of variables but polynomial in their cardinality. Hence, even though an allele BN with more than three individuals and/or more than one gene leads to a non-chordal factor graph with a larger tree-width, they usually lower time complexities for inferences. Let $\mathcal{G} = \{G^1, \ldots, G^N\}$ be a set of N genes segregating independently, an allele BN is defined as $\mathcal{B}^{\text{allele}} = \left(G^{\text{allele}} = \left(\{\{A^{g,p}\}_{g=1,\ldots,N}, \{A^{g,m}\}_{g=1,\ldots,N}, Z\}, \mathcal{E}^{\text{allele}}\right), \mathbb{P}\right)$ where, for all $g \in \{1, \ldots, N\}$, $A^{g,p} = \{A_i^{g,p}\}_{i \in \{1,\ldots,n\}}$ (respectively $A^{g,m} = \{A_i^{g,m}\}_{i \in \{1,\ldots,n\}}$) is the set of paternal (respectively maternal) alleles for gene G^g, $Z = \{Z_i\}_{i \in \{1,\ldots,n\}}$ is the set of phenotypes and $\mathcal{E}^{\text{allele}} = \{\{(A_{\mathrm{p}(i)}^{g,p}, A_i^{g,p})_{i \in \overline{\mathcal{F}}}\}_{g \in \{1,\ldots,N\}}, \{(A_{\mathrm{p}(i)}^{g,m}, A_i^{g,p})_{i \in \overline{\mathcal{F}}}\}_{g \in 1,\ldots,N}$, $\{(A_{\mathrm{m}(i)}^{g,p}, A_i^{g,m})_{i \in \overline{\mathcal{F}}}\}_{g \in \{1,\ldots,N\}}, \{(A_{\mathrm{m}(i)}^{g,m}, A_i^{g,m})_{i \in \overline{\mathcal{F}}}\}_{g=1,\ldots,N}, \{(A_i^{g,p}, Z_i)_{i \in \{1,\ldots,n\}}\}_{g \in 1,\ldots,N}$, $\{(A_i^{g,m}, Z_i)_{i \in \{1,\ldots,n\}}\}_{g \in 1,\ldots,N}\}$. A graphical representation of the DAG associated with an allele BN for the pedigree represented in Figure 3.7 is proposed in Figure 3.10 in the particular case of a single gene G^1. The joint probability of $\{A^{g,p}\}_{g=1,\ldots,N}$, $\{A^{g,m}\}_{g=1,\ldots,N}$ and Z is given by the following product of CPDs:

$$\begin{aligned}&\mathbb{P}\left(\{A^{g,p}\}_{g=1,\ldots,N}, \{A^{g,m}\}_{g=1,\ldots,N}, Z | S, \theta\right) = \\ &\qquad \prod_{i=1}^{n} \left\{ \prod_{g=1}^{N} \mathbb{P}(A_i^{g,p} | A_{\mathrm{p}(i)}^{g,p}, A_{\mathrm{p}(i)}^{g,m}; \theta^x) \mathbb{P}(A_i^{g,m} | A_{\mathrm{m}(i)}^{g,p}, A_{\mathrm{m}(i)}^{g,m}; \theta^x) \right\} \\ &\qquad\qquad \times \mathbb{P}(Z_i | \{A_i^{g,p}\}_{g=1,\ldots,N}, \{A_i^{g,m}\}_{g=1,\ldots,N}, S_i; \theta^z) \end{aligned} \tag{3.5}$$

where $\theta = \{\theta^x, \theta^z\}$, for all $i \in \mathcal{F}$, $\mathrm{p}(i) = \mathrm{m}(i) = 0$ and for all $g \in \{1, \ldots, N\}$, $A_0^{g,p} = A_0^{g,m} = \emptyset$ by convention.

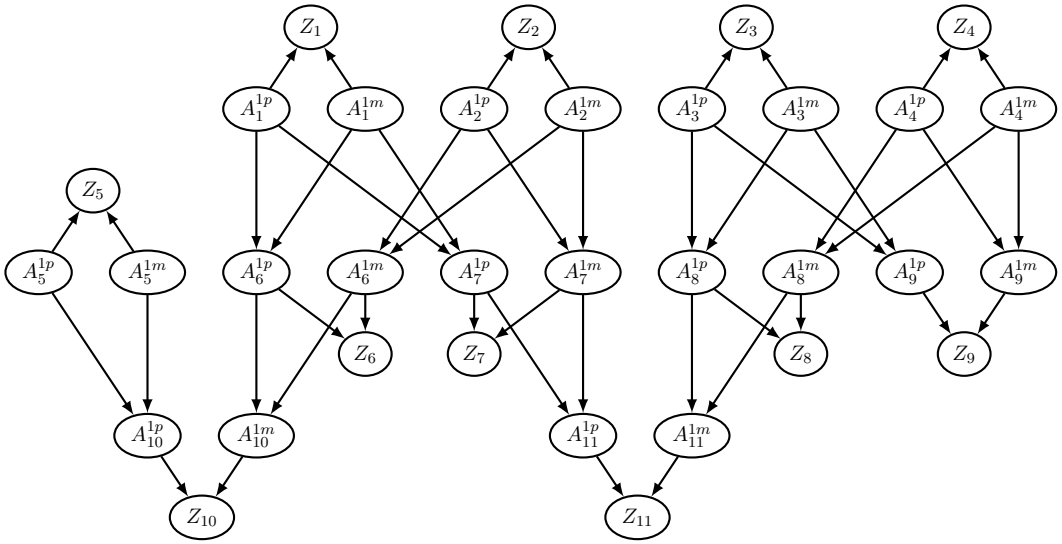


Figure 3.10: Allele DAG associated with the pedigree represented Figure 3.7 in the framework of a single gene G^1 whose paternal (respectively maternal) component allele carried by individual $i \in \{1, \ldots, n\}$ is denoted $A_i^{1,p}$ (respectively $A_i^{1,m}$).

	$A^{g,p}_{\mathrm{p}(i)} = 0,$ $A^{g,m}_{\mathrm{p}(i)} = 0$	$A^{g,p}_{\mathrm{p}(i)} = 1,$ $A^{g,m}_{\mathrm{p}(i)} = 0$	$A^{g,p}_{\mathrm{p}(i)} = 0,$ $A^{g,m}_{\mathrm{p}(i)} = 1$	$A^{g,p}_{\mathrm{p}(i)} = 1,$ $A^{g,m}_{\mathrm{p}(i)} = 1$
$A_i^{g,p} = 0$	1.0	0.5	0.5	0.0
$A_i^{g,p} = 1$	0.0	0.5	0.5	1.0

Table 3.2: CPDs $\mathbb{P}(A_i^{g,p}|A^{g,p}_{\mathrm{p}(i)}, A^{g,m}_{\mathrm{p}(i)}, \theta^x)$ for non-founders in an allele BN where $A_i^{g,p}$ (respectively $A_i^{g,m}$) denote the paternal (respectively maternal) allele for gene G^g (assumed to be biallelic) carried by individual $i \in \{1, \ldots, n\}$.

genotypic component Under Assumption A1 each variable $A_i^{g,p}$ and $A_i^{g,m}$ takes value 0 is non-pathogenic and 1 otherwise. Under Assumptions A1, A2 and A3, the parameter θ^x is restricted to $\theta^x = q = (q_1, \ldots, q_N)$, the vector of pathogenic allele frequencies, where, for $g \in \{1, \ldots, N\}$, q_g is the frequency of pathogenic alleles for gene G^g in the general population. Under Assumption A2, for all $i \in \mathcal{F}$, $g \in \{1, \ldots, N\}$ and $h \in \{p, m\}$, $\mathbb{P}(A_i^{g,h} = 0|q) = (1 - q_g)$ and $\mathbb{P}(A_i^{g,h} = 1|q) = q_g$. Under Assuption A3, for $i \in \overline{\mathcal{F}}$ and $g \in \{1, \ldots, N\}$, $\mathbb{P}(A_i^{g,p}|A^{g,p}_{\mathrm{p}(i)}, A^{g,m}_{\mathrm{p}(i)})$ (and similarly for $\mathbb{P}(A_i^{g,m}|A^{g,p}_{\mathrm{m}(i)}, A^{g,m}_{\mathrm{m}(i)})$) are given in Table 3.2.

Phenotypic component As $X_i^g = A_i^{g,p} A_i^{g,m}$ and $X_i = (X_i^g)_{g=1,\ldots,N}$, CPDs of the form $\mathbb{P}(Z_i|\{A_i^{g,p}\}_{g=1,\ldots,N}, \{A_i^{g,m}\}_{g=1,\ldots,N}, S_i; \theta^z)$ are similarly defined as those of a genotype BN (see Equation (3.2) and Section 3.2.2.1) where each genotype is split into its component alleles.

3.2.2.3 Selector Bayesian network

When several genes are in linkage disequilibrium, their non-independent segregation (also called haplotype information) is usually taken into account by adding binary variables called selectors who denote the (paternal or maternal) origin of each allele of the offsprings. Selector BNs were introduced by Kong et al. (1991) and Thompson (1994). Additional edges between appropriate alleles in an allele BN could be considered but a selector BN is usually preferred as it allows for decreasing the state spaces of CPDs. A selector BN is defined as $\mathcal{B}^{\text{sel}} = \big(G^{\text{sel}} = (\{\{A^{g,p}\}_{g\in\{1,\dots,N\}}, \{A^{g,m}\}_{g\in\{1,\dots,N\}}, \{\text{SA}^{g,p}\}_{g\in\{1,\dots,N\}}, \{\text{SA}^{g,m}\}_{g\in\{1,\dots,N\}}, Z\}, \mathcal{E}^{\text{sel}}), \mathbb{P}\big)$ where, for $g \in \{1,\dots,N\}$ and $h \in \{p,m\}$, $\text{SA}^{g,h} = \{\text{SA}_i^{g,h}\}_{i\in\overline{\mathcal{F}}}$ such that $\text{SA}_i^{g,h}$ denotes the selector of allele $A_i^{g,h}$. A selector is a binary variable and takes value p (respectively m) if the corresponding allele comes from the paternal (respectively maternal) chromosome of the parent of origin. For the sake of simplicity, we restrict this section to two genes G^1 and G^2. The non-independent segregation of both genes is taken into account with an edge between $\text{SA}_i^{1,h}$ and $\text{SA}_i^{2,h}$ for all $h \in \{p,m\}$ and $i \in \overline{\mathcal{F}}$. The DAG of a selector BN associated with a simple trio composed of a father, a mother and an offspring respectively indexed by 1, 2 and 3 and two genes G^1, G^2 is represented in Figure 3.11.

The joint probability of $\{A^{g,p}\}_{g\in\{1,2\}}, \{A^{g,m}\}_{g\in\{1,2\}}, \{\text{SA}^{g,p}\}_{g\in\{1,2\}}, \{\text{SA}^{g,m}\}_{g\in\{1,2\}}$ and Z is given by the following product of CPDs:

$$\begin{aligned}
&\mathbb{P}\left(\{A^{g,p}\}_{g\in\{1,2\}}, \{A^{g,m}\}_{g\in\{1,2\}}, \{\text{SA}^{g,p}\}_{g\in\{1,2\}}, \{\text{SA}^{g,m}\}_{g\in\{1,2\}}, Z|S, \theta^{\text{sel}}\right) = \\
&\quad\prod_{i=1}^{n}\left\{\prod_{g\in\{1,2\}}\mathbb{P}(A_i^{g,p}|A_{\text{p}(i)}^{g,p}, A_{\text{p}(i)}^{g,m}, \text{SA}_i^{g,p})\mathbb{P}(A_i^{g,m}|A_{\text{m}(i)}^{g,p}, A_{\text{m}(i)}^{g,m}, \text{SA}_i^{g,m})\right\}^{\mathbb{1}_{\{i\in\overline{\mathcal{F}}\}}} \\
&\qquad\times\left\{\prod_{h\in\{p,m\}}\mathbb{P}(\text{SA}_i^{1,h})\mathbb{P}(\text{SA}_i^{2,h}|\text{SA}_i^{1,h};\delta)\right\}^{\mathbb{1}_{\{i\in\overline{\mathcal{F}}\}}} \\
&\qquad\quad\times\left\{\prod_{g\in\{1,2\}}\mathbb{P}(A_i^{g,p}|q)\mathbb{P}(A_i^{g,m}|q)\right\}^{\mathbb{1}_{\{i\in\mathcal{F}\}}} \\
&\qquad\qquad\times\mathbb{P}\left(Z_i|\{A_i^{g,p}\}_{g\in\{1,2\}}, \{A_i^{g,m}\}_{g\in\{1,2\}}, S_i;\theta^z\right) \quad (3.6)
\end{aligned}$$

where $\theta^{\text{sel}} = \{q, \delta, \theta^z\}$ such that δ is the genetic distance (Haldane, 1919), in Morgan units, between genes G^1 and G^2.

We assume a Mendelian inheritance for G^1 and therefore, for $i \in \overline{\mathcal{F}}$ and $h \in \{p,m\}$, $\mathbb{P}(\text{SA}_i^{1,h})$ is uniform over $\{0,1\}$. Recalling Equation (3.1) which links the genetic distance between two loci and the probability of an odd number of crossing-overs between them during gametogenesis, under the hypothesis that the number of crossing-overs between the two loci follows a Poisson distribution of parameter δ, for all $h \in \{p,m\}$, CPDs of the form $\mathbb{P}(\text{SA}_i^{2,h}|\text{SA}_i^{1,h};\delta)$ are given, for $s \in \{p,m\}$, by

$$\mathbb{P}(\text{SA}_i^{2,h} = s|\text{SA}_i^{1,h} = s;\delta) = \frac{1-e^{-2\delta}}{2} \quad \text{and} \quad \mathbb{P}(\text{SA}_i^{2,h} \neq s|\text{SA}_i^{1,h} = s;\delta) = 1-\frac{1-e^{-2\delta}}{2}.$$

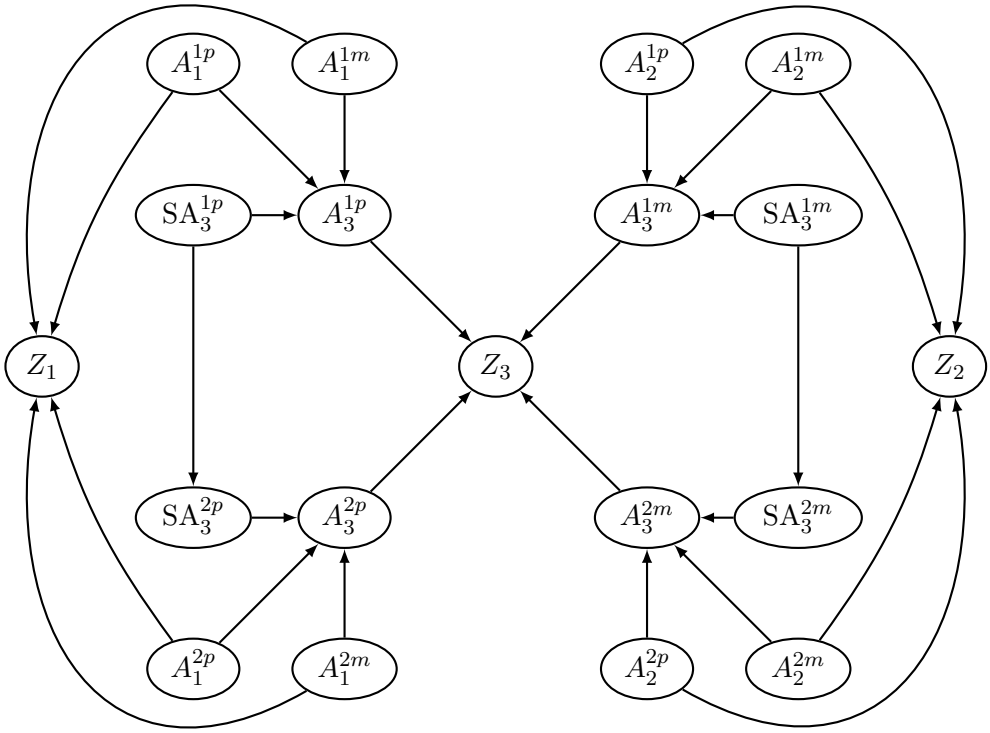


Figure 3.11: Selector DAG associated with a simple trio including a father, a mother and an offspring respectively indexed by 1, 2, and 3, and two genes G^1 and G^2 such that, for $g \in \{1, 2\}$, the component paternal (respectively maternal) allele of gene G^g carried by individual $i \in \{1, 2, 3\}$ is denoted $A_i^{g,p}$ (respectively $A_i^{g,m}$).

Finally, for $i \in \overline{\mathcal{F}}$, for $g \in \{1,2\}$, CPDs of the form $\mathbb{P}(A_i^{g,p}|A_{\mathrm{p}(i)}^{g,p}, A_{\mathrm{p}(i)}^{g,m}, \mathrm{SA}_i^{g,p})$ (and similarly for $\mathbb{P}(A_i^{g,m}|A_{\mathrm{m}(i)}^{g,p}, A_{\mathrm{m}(i)}^{g,m}, \mathrm{SA}_i^{g,m})$) are given, for $a,b \in \{0,1\}$, by $\mathbb{P}(A_i^{g,p} = a|A_{\mathrm{p}(i)}^{g,p} = a, A_{\mathrm{p}(i)}^{g,m} = b, \mathrm{SA}_i^{g,p} = p) = 1$, $\mathbb{P}(A_i^{g,p} = a|A_{\mathrm{p}(i)}^{g,p} = b, A_{\mathrm{p}(i)}^{g,m} = a, \mathrm{SA}_i^{g,p} = m) = 1$.

3.2.3 Inferences and main algorithms

As a pedigree is a BN, many intractable computations using brute force become tractable with the sum-product algorithm detailed in Chapter 1. We define the evidence ev to be a set of observed data, usually a set of phenotypes and/or results of germline sequencing. There are two main domains involving complex computations:

- **Parameter estimation in genetic epidemiology.** Parameter estimation usually involves MLE methods to maximize the likelihood of the model given by $\mathcal{L}(\theta) = \mathbb{P}(\mathrm{ev}|\theta)$ where the parameter θ is $\theta = \{\theta^x, \theta^z\}$ in segregation analysis or $\theta = \delta$ in linkage analysis (see Section 3.1.3.2 for details). In multipoint linkage analysis, with an increasing number of markers, the likelihood of the model may become intractable and many authors proposed approximate inferences and/or sampling methods. Note that when an exact computation is tractable, the likelihood can be computed with an inward pass of the sum-product algorithm choosing any variable as a root.

- **Risks of genetic predisposition computation.** For a fixed parameter, one computes, for a family member i (or a set of family members), the probability of carrying a pathogenic allele involved in a disease of interest conditional on ev. Conditional probabilities of the form $\mathbb{P}(X_i|\mathrm{ev})$ in a genotype BN and of the form $\mathbb{P}(\{A_i^{g,p}\}_{g\subseteq\{1,\dots,N\}}, \{A_i^{g,m}\}_{g\subseteq\{1,\dots,N\}}|\mathrm{ev})$ in an allele or a selector BN can be computed with an inward pass of the sum-product algorithm choosing, as the root, a clique containing the variables of interest, $\{X_i\}$ or $\{A_i^{g,p}, A_i^{g,m}\}_{g\subseteq\{1,\dots,N\}}$, according to the chosen model. Note that there always exists at least one such clique as $\{A_i^{g,p}, A_i^{g,m}\}_{g=\{1,\dots,N\}}$ are common (graph) parents of Z_i, thus there exists a potential containing all of them in its scope. Performing an additional outward pass for the same time complexity allows one for computing risks of genetic predisposition for any other member in the family (see Theorems 2 and 3).

Several computational shortcuts detailed in Section 1.6 are applicable in pedigree-based models such as:

- **Reduction of potentials.** Reduce CPDs containing variables assigned a hard evidence. For instance, if Z_i is assigned a hard evidence, the CPD of the form $\mathbb{P}(Z_i|\{A_i^{g,p}\}_{g=1,\dots,N}, \{A_i^{g,m}\}_{g=1,\dots,N})$ can be reduced, after entering the evidence, to the potential $\phi_{Z_i}(\{A_i^{g,p}\}_{g=1,\dots,N}, \{A_i^{g,m}\}_{g=1,\dots,N})$ leading to the removal of variable Z_i in the factor graph.

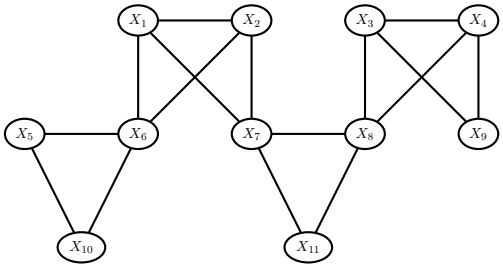
- **Pruning.** Remove unobserved variables with unobserved (graph) descendent if no latter inference for that variable is needed, i.e. ignore its CPD. This is

particularly obvious for an unobserved phenotype Z_i. Furthermore if individual i has no descendent, pruning Z_i allows also for avoiding edges $(A_i^{g,p}, A_i^{g,m})$ for $g \in \{1, \ldots, N\}$ and $(A_i^{g_1,h_1}, A_i^{g_2,h_2})$ for $g_1, g_2 \in \{1, \ldots, N\}$ and $h_1, h_2 \in \{1, 2\}$ in the resulting factor graph.

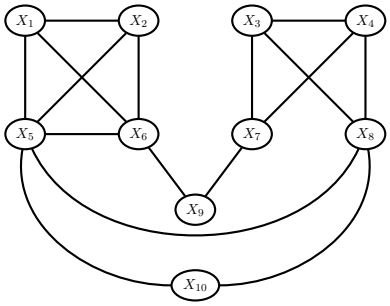
- **Forcing.** Assign a hard evidence whenever possible. As most variables in pedigree-based models have deterministic relationships, the introduction of an evidence on genotypes may lead to other possible hard assignments. For instance assume that genotype $X_1^1 = 11$ is observed in the DAG represented in Figure 3.10, we have $A_1^{1,p} = 1$ and $A_1^{1,m} = 1$, hence $A_6^{1,p}$ is deterministically equal to 1 and the event $\{A_6^{1,p} = 1\}$ can be added to the evidence leading to state space(s) reduction and/or a sparser factor graph. Forcing is particularly efficient for linkage analysis with possibly many forcing options. An example of complexity reduction obtained at each computational shortcut is proposed in Table 5.3 (Chapter 5) in a simple framework of two-point linkage over a dataset extracted from the Mendel package (Lange et al., 2013).

As phenotypes are, in most situations, either unobserved and removed or assigned a hard evidence, they usually do not appear in scopes of potentials nor in factor graphs. We propose in Figure 3.12 a graphical representation of factor graphs associated with different DAGs where the evidence is the set of observed phenotypes. Note that a genotype DAG with conventional mating lead to a chordal factor graph (Figure 3.12a) but complex mating such as consanguinity or cross mating lead to non-chordal factor graphs containing several cycles such as $X_5 - X_1 - X_6 - X_9 - X_7 - X_3 - X_8 - X_5$ among others in Figure 3.12b. Such cycles in genetics are called *mating loops*, not to be confounded with loops in graph theory. Finally an allele DAG with more than three individuals and a selector DAG lead to non-chordal factor graphs. Finding good elimination orderings for triangulating such non-chordal graphs, especially those including selectors, is one of the main issues for estimations or inferences in pedigree-based models.

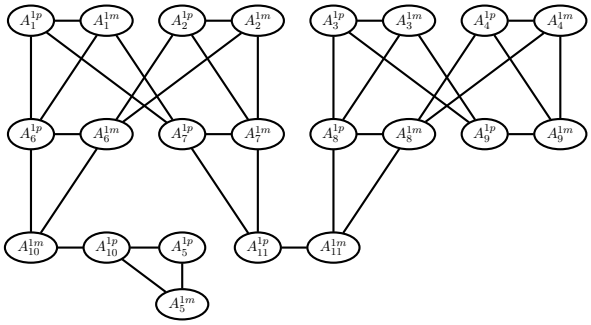
Complexity reduction to perform computations in pedigree-based models has been studied for decades in genetics before and/or simultaneously with probabilistic graphical models theory, starting with the Elston-Stewart (ES) algorithm published in 1971 (Elston and Stewart, 1971) based on an elimination of genotypes from last generation towards first generation, hence an inward pass in a genotype BN, called *peeling* in genetics. The ES algorithm has been followed by a vast literature proposing several extensions of the algorithm. For instance Elston (1973) extended the algorithm for time-to-event phenotypes as well as for correcting the bias induced by ascertainment. Indeed families are studied conditional on the fact that the proband came for investigation. In 1974, Ott (1974) extended the ES algorithm to two-point linkage analysis. In 1975 Lange and Elston (1975) proposed a method based on pedigree duplication in order to reduce computational complexity in pedigrees with mating loops. Their article was soon followed by a new method of lower time complexity for inferences in pedigrees with mating loops developed by Cannings et al. (1976). The authors introduce the notion of *cutsets* and create functions of a set of remaining variables during peeling which deeply refers to fill-in edges and triangu-



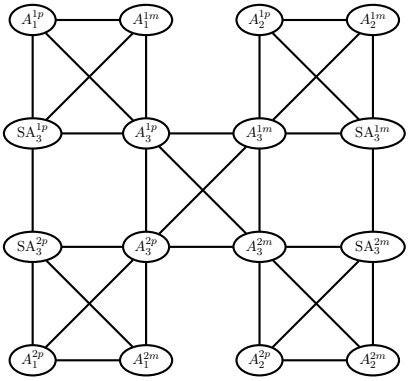
(a) Factor graph associated with the genotype DAG represented in Figure 3.9a



(b) Factor graph associated with the genotype DAG represented Figure 3.9b



(c) Factor graph associated with the allele DAG represented Figure 3.10



(d) Factor graph associated with the selector DAG represented Figure 3.11

Figure 3.12: Factor graphs associated with various DAGs commonly used in genetic analyses where all genotypes are latent variables and the evidence is the set of observed phenotypes. Note several cycles in Figure 3.12b (for instance $X_5 - X_1 - X_6 - X_9 - X_7 - X_3 - X_8 - X_5$) induced by the complex mating in the pedigree represented in Figure 3.8.

lation (See also Totir et al., 2009, for a description of a general framework linking exact inference in probabilistic graphical models and pedigree-analysis).

Multipoint linkage analysis addresses the problem of mapping genes on the genome. Computations involve selector BNs with multiple markers. Because of the structure of the factor graph, the ES algorithm becomes rapidly intractable when considering more than a few markers. In order to handle multiple markers, the Lander-Green (LG) algorithm published in (Lander and Green, 1987) and originally implemented in the MAPMAKER computer program (Lander et al., 1987) is based on peeling loci one after the other. Whereas the ES algorithm is well adapted for pedigrees of arbitrary sizes and only few loci, the LG algorithm is appropriate for many dependent loci and pedigrees of limited size. The complex structure of factor graphs in multiple linkage has favor intensive research in probabilistic graphical models in this domain. The LG algorithm performs exact computations, however such computations become intractable with an increasing number of loci. Most methods in multipoint genetic linkage are based on approximations (Kong et al., 1991), sampling methods (Thompson and Heath, 1999; Wijsman et al., 2006) or a combination of exact methods and sampling (Jensen and Kong, 1999) (See also Thompson, 2000, for a details on different methods in pedigree analysis). Numerous algorithms have been developed and are regularly updated for multipoint linkage analyses among which (along with their original publication) FASTLINK (Cottingham Jr et al., 1993; Becker et al., 1998), VITESSE (O'Connell and Weeks, 1995), GENEHUNTER (Kruglyak et al., 1996), SimWalk2 (Sobel and Lange, 1996; Sobel et al., 2001, 2002), ALLEGRO (Gudbjartsson et al., 2000), SUPERLINK (Fishelson and Geiger, 2002) and its parallelized version (Silberstein et al., 2006).

Part II

Belief propagation in polynomials

Chapter 4

Introduction to generating functions in Bayesian networks

Sommaire

4.1 Introduction

All examples chosen for illustrating the sum-product algorithm in Chapter 1 are based on real-valued potentials. However, one can think of various mathematical objects with the right properties towards the sum and product operator in order to compute other quantities of interest. For instance the distribution and moments of the number of events has been extensively studied, in particular in Markov chains and HMMs, through the computation of generating functions. In this introductory chapter we review such methods before exploring other potential uses of generating functions in HMMs and BNs in Chapters 5, 6 and 8 in which we respectively study the derivatives of the likelihood in BNs, the number of segments of similar composition in sequences and the number of genotypes carrying deleterious mutations among family members in pedigree-based models.

An active domain of interest is the detection of patterns of unexpected frequency in a sequence with applications in various fields and in particular in molecular biology since the 90's. In classical statistical models, letters are assumed to be generated by a Bernoulli or a Markov model over a finite alphabet (Robin and Daudin, 1999; Régnier, 2000; Robin et al., 2005). The exact distribution of the number of a pattern, also called a word, is given by the coefficients of the Taylor expansion of its probability generating function (pgf) which is a rational function (Régnier, 2000). In practice,

for long sequences, an exact inference is intractable and approximate distribution can be obtained with a Gaussian approximation (Prum et al., 1995), a compound Poisson approximation (Arratia et al., 1990; Geske et al., 1995; Schbath, 1997) or large deviations (Fu et al., 2003; Nuel, 2004). A review of these methods is proposed by Reinert et al. (2000) and Lothaire (2005).

Turning the problem into the study of a regular expression via deterministic finite automata (DFA), Nicodeme et al. (2002) allow for studying a broader range of patterns and an exact inference of the distribution and the moment generating function (mgf) of the counts in most cases. Several authors extended this idea to a Markov chain embedding of pattern occurrences (Fu and Koutras, 1994; Lladser; Nuel, 2008, 2010; Nuel et al., 2010). The sum-product algorithm is exploited to reduce computational complexity.

The distribution and moments of the number of pattern in heterogeneous Markov models via HMMs has later been studied by Aston and Martin (2007); Martin and Aston (2013); Nuel (2019). In Nuel (2019), the authors also allow for studying a broader type of patterns through the computation of conditional pgf and mgf in constrained HMM. Constraints are applied by entering an *evidence* of interest in the HMM.

Replacing the sum-product algorithm in HMMs by the sum-product algorithm in BNs, the extension of the methods developed in HMMs to BNs is straightforward as previously mentioned by Martin and Aston (2013) for pgf and Cowell (1992); Nilsson (2001) for mgf. In this introductory chapter, we review these methods respectively in Section 4.2 and 4.3 for pgf and mgf using notation introduced in Chapter 1 and propose an illustrative toy example in Section 4.4. For guiding the reading throughout the manuscript, a polynomial potential or a functions taking its values in the polynomial ring will be bolded.

4.2 Distribution of the number of events

Let $\mathcal{B} = (G = (X = \{X_1, \ldots, X_n\}, \mathcal{E}), \mathbb{P})$ be a BN such that, for all $u \in \{1, \ldots, n\}$, X_u is a discrete variable taking its values in $\mathcal{X}_u$ with $|\mathcal{X}_u| < \infty$. For any subset $U \subseteq \{1, \ldots, n\}$, we denote the set $\{X_u\}_{u \in U}$ by X_U. By definition of a BN, the joint probability of $X_1, \ldots, X_n$ factorizes according to the following product of Conditional Probability Distributions (CPDs):

$$\mathbb{P}(X_1, \ldots, X_n) = \prod_{u=1}^{n} \mathbb{P}(X_u | X_{\mathrm{pa}(u)})$$

where $\mathrm{pa}(u)$ is the set of labels for graph parents of X_u in the directed acyclic graph G. Let U be a subset of $\{1, \ldots, n\}$, we introduce

$$N = \sum_{u \in U} N_u \quad \text{with} \quad N_u = \mathbb{1}_{\{X_{\bar{u}} \in \mathcal{Y}_{\bar{u}} \subset \mathcal{X}_{\bar{u}}\}}$$

where $\mathbb{1}_{\{.\}}$ is the indicator function, $\bar{u} = (\mathrm{pa}(u), u)$ in topological ordering and $\mathcal{X}_{\bar{u}}$ is the set of values taken by $X_{\bar{u}}$. In practice N_u could be any function of $X_{\{\mathrm{pa}(u),u\}}$

or a subset of $X_{\{\mathrm{pa}(u),u\}}$ but, for the sake of simplicity, we restrict explanations to the above definition. A simple example is for instance given over a binary BN where each variable takes its values in $\{0,1\}$. Assume that we are interested in the number N of variables taking value 1, then, for all $u \in \{1,\dots,n\}$, sets $\mathcal{Y}_{\bar{u}}$ are defined as $\mathcal{Y}_{\bar{u}} = \{0,1\}^{|X_{\mathrm{pa}(u)}|} \times 1$.

For each $u \in \{1,\dots,n\}$, we introduce the potential, polynomial in z,

$$\boldsymbol{\xi}_u(X_{\{\mathrm{pa}(u),u\}}) = \mathbb{P}(X_u|X_{\mathrm{pa}(u)})z^{\mathbb{1}_{\{u\in U\}}N_u}, \tag{4.1}$$

leading to the following expression of the probability generating function of N, $\mathrm{pgf}_N(z) = \mathbb{E}\big[z^N\big]$:

$$\mathrm{pgf}_N(z) = \sum_{k=0}^{\infty} \mathbb{P}(N=k)z^k = \sum_X \prod_{u=1}^{n} \boldsymbol{\xi}_u(X_{\{\mathrm{pa}(u),u\}}). \tag{4.2}$$

Indeed, since each value taken by $\boldsymbol{\xi}_u(X_{\{\mathrm{pa}(u),u\}})$ is a polynomial of degree N_u (or zero if $u \notin U$), we keep track of the count and then we have to sum over all possible configurations of X to get $\sum_X \sum_{k=0}^{\infty} \mathbb{P}(X, N=k)z^k$.

Entering evidence. Let $\mathrm{ev} = \cap_{u\in E}\{X_u \in \mathcal{X}_u^* \subset \mathcal{X}_u\}$ be an evidence for $\mathcal{B}$ where $E \subset \{1,\dots,n\}$ and for all $u \notin E$, let $\mathcal{X}_u^* = \mathcal{X}_u$, we define the conditional pgf of N to be $\mathrm{pgf}_{N|\mathrm{ev}}(z) = \mathbb{E}\big[z^N|\mathrm{ev}\big]$ which is given by

$$\mathrm{pgf}_{N|\mathrm{ev}}(z) = \sum_{k=0}^{\infty} \mathbb{P}(N=k|\mathrm{ev})z^k = \frac{1}{\mathbb{P}(\mathrm{ev})} \sum_X \prod_{u=1}^{n} \boldsymbol{f}_u(X_{\{\mathrm{pa}(u),u\}}) \tag{4.3}$$

where, for all $u \in \{1,\dots,n\}$,

$$\boldsymbol{f}_u(X_{\{\mathrm{pa}(u),u\}}) = \mathbb{1}_{\{X_v\in\mathcal{X}_v^*,\forall v\in\{u,\mathrm{pa}(u)\}\}}\mathbb{P}(X_u|X_{\mathrm{pa}(u)})z^{\mathbb{1}_{\{u\in U\}}N_u}.$$

Let $\boldsymbol{\psi}(X) = \prod_{u=1}^{n} \boldsymbol{f}_u(X_{\{\mathrm{pa}(u),u\}})$, we restrict our work to discrete variables of finite cardinality, hence $\sum_X \boldsymbol{\psi}(X)$ is of finite degree and $\mathbb{P}(\mathrm{ev}) = \sum_N \sum_X \mathbb{P}(N, X, \mathrm{ev})$ is given by the sum of all coefficients in $\sum_X \boldsymbol{\psi}(X)$.

Practical computation. The time complexity of a brute force computation of Equation (4.2) or (4.3) is in $\mathcal{O}\left(K \times \prod_{u=1}^{n} |\mathcal{X}_u|\right)$ where K is the maximal value taken by N. However, one can apply an upward pass of the sum-product algorithm detailed in Algorithm 4 for a complexity reduction. We briefly adapt, in this section, quantities seen in Chapter 1 for the computation of the conditional pgf of N and the extension to the pgf of N is straightforward by replacing $\boldsymbol{f}$ and $\boldsymbol{f}_u$ by $\boldsymbol{\xi}$ and $\boldsymbol{\xi}_u$ in this paragraph. Let $\boldsymbol{f}$ be the set of potentials $\{\boldsymbol{f}_u\}_{u=1,\dots,n}$ and σ be an elimination ordering over X, let $J = (C = \{C_1,\dots,C_m\}, \mathcal{F})$ be a junction-tree defined by $\mathrm{VE}(\boldsymbol{f}, X, \sigma)$, we define, for all $i \in \{1,\dots,m\}$, the clique potential of C_i denoted $\boldsymbol{\Phi}_i$ as in Equation (1.9). Let C_m be a randomly chosen root clique and $\boldsymbol{\delta} = \{\boldsymbol{\delta}_i\}_{i=1,\dots,m}$ be the set of messages $\boldsymbol{\delta}_{i\to\mathrm{to}(i)}$ as defined in Equation (9), one can recursively compute the set $\boldsymbol{\delta}$ using Corollary 1 with an upward pass of Algorithm 4

in $\mathcal{O}\left(K \times \sum_{i=1}^{m} \prod_{X_u \in C_i} \max_u |\mathcal{X}_u|\right)$ time complexity, i.e, if all variables are of similar cardinality, $\mathcal{O}\left(K \times m \times \max_u |\mathcal{X}_u|^{\tau+1}\right)$ where τ is the tree-width of J. At the end of the recursion, one can obtained, by a direct application of Theorem 3

$$\sum_X \boldsymbol{\psi}(X) = \sum_{C_m} \boldsymbol{\psi}(C_m) = \sum_{C_m} \boldsymbol{\Phi}_m(C_m) \prod_{i, i-m} \boldsymbol{\delta}_{i \to m}(S_{i,m}). \tag{4.4}$$

Remark. One can choose an arbitrary maximal number of events $k_{\max} < K$ and replace, in the algorithm, the conventional product of potentials by a product with an additional truncation at the order $k_{\max}$, the complexity of the upward pass to compute the distribution or conditional distribution of N from $N = 0$ up to $N = k_{\max}$ becomes $\mathcal{O}\left(k_{\max} \times \sum_{i=1}^{m} \prod_{X_u \in C_i} \max_u |\mathcal{X}_u|\right)$ in time.

Extension to the joint distribution of multiple types of events. Extending Equations (4.2) and (4.3) to the joint distribution of the number of multiple types of events is straightforward with multiple dummy variables and multivariate polynomials but increases computational cost. Let U and U' be two subsets of $\{1, \ldots, n\}$, let $N = \sum_{u \in U} N_u$ and $M = \sum_{u \in U'} M_u$ where, for all $u \in U$ (respectively $u \in U'$), $N_u = \mathbb{1}_{\{\cap_{v \in \{\mathrm{pa}(u), u\}} X_v \in \mathcal{Y}_v \subseteq \mathcal{X}_v\}}$ (respectively $M_u = \mathbb{1}_{\{\cap_{v \in \{\mathrm{pa}(u), u\}} X_v \in \mathcal{Z}_v \subseteq \mathcal{X}_v\}}$), we define for each $u \in \{1, \ldots, n\}$,

$$\widetilde{\boldsymbol{\xi}}_u(X_{\{\mathrm{pa}(u), u\}}) = \mathbb{P}(X_u | X_{\mathrm{pa}(u)}) y^{\mathbb{1}_{\{u \in U\}} N_u} z^{\mathbb{1}_{\{u \in U'\}} M_u} \tag{4.5}$$

and

$$\widetilde{\boldsymbol{f}}_u(X_{\{\mathrm{pa}(u), u\}}) = \mathbb{1}_{\{X_v \in \mathcal{X}_v^*, \forall v \in \{u, \mathrm{pa}(u)\}\}} \mathbb{P}(X_u | X_{\mathrm{pa}(u)}) y^{N_u} z^{M_u} \tag{4.6}$$

where y and z are two dummy variables. We have

$$\sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} \mathbb{P}(N = k, M = \ell) y^k z^\ell = \sum_X \prod_{u=1}^{n} \widetilde{\boldsymbol{\xi}}_u(X_{\{\mathrm{pa}(u), u\}}) \tag{4.7}$$

and

$$\sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} \mathbb{P}(N = k, M = \ell | \mathrm{ev}) y^k z^\ell \propto \sum_X \prod_{u=1}^{n} \widetilde{\boldsymbol{f}}_u(X_{\{\mathrm{pa}(u), u\}}). \tag{4.8}$$

Similarly one can compute Equations (4.7) and (4.8) with an upward pass of the sum-product algorithm in $\mathcal{O}\left(K \times L \times m \times \max_u |\mathcal{X}_u|^{\tau+1}\right)$ in time where K (respectively L) is the maximal value taken by N (respectively M). One can arbitrarily choose $k_{\max} < \max(K, L)$ in order to reduce the complexity to $\mathcal{O}\big(\min(K, L) \times k_{\max} \times m \times \max_u |\mathcal{X}_u|^{\tau+1}\big)$ if $k_{\max} \geq \min(K, L)$ or $\mathcal{O}\big(k_{\max}^2 \times m \times \max_u |\mathcal{X}_u|^{\tau+1}\big)$ otherwise.

4.3 Moments of the number of events

With a very similar reasoning, the moment generating function of N is given by marginalizing X in a product of potentials, in brief, by replacing z by e^t in quantities

introduced in the previous section (Cowell, 1992; Nilsson, 2001). The mgf of N, $\mathrm{mgf}_N(t) = \mathbb{E}\left[e^{tN}\right]$ can be expressed by

$$\mathrm{mgf}_N(t) = \sum_{k=0}^{\infty} \mathbb{P}(N=k)e^{tk} = \sum_X \prod_{u=1}^{n} \boldsymbol{\zeta}_u(X_{\{\mathrm{pa}(u),u\}}) \tag{4.9}$$

where

$$\boldsymbol{\zeta}_u(X_{\{\mathrm{pa}(u),u\}}) = \mathbb{P}(X_u|X_{\mathrm{pa}(u)})e^{t\times \mathbb{1}_{u\in U} N_u}. \tag{4.10}$$

We define the conditional mgf of N to be $\mathrm{mgf}_{N|\mathrm{ev}}(t) = \mathbb{E}\left[e^{tN}|\mathrm{ev}\right]$ which is given by

$$\mathrm{mgf}_{N|\mathrm{ev}}(t) = \frac{1}{\mathbb{P}(\mathrm{ev})} \sum_{k=0}^{\infty} \mathbb{P}(N=k|\mathrm{ev})e^{tk} = \frac{1}{\mathbb{P}(\mathrm{ev})} \sum_X \prod_{u=1}^{n} \boldsymbol{g}_u(X_{\{\mathrm{pa}(u),u\}}) \tag{4.11}$$

where

$$\boldsymbol{g}_u(X_{\{\mathrm{pa}(u),u\}}) = \mathbb{1}_{\{X_v\in\mathcal{X}_v^*,\forall v\in\{u,\mathrm{pa}(u)\}\}}\mathbb{P}(X_u|X_{\mathrm{pa}(u)})e^{t\times \mathbb{1}_{u\in U} N_u} \tag{4.12}$$

and $\mathbb{P}(\mathrm{ev})$ is given by the first coefficient in $\sum_X \prod_{u=1}^{n} \boldsymbol{g}_u(X_{\{\mathrm{pa}(u),u\}})$.

In practice, one can implement the potentials $\boldsymbol{\zeta}_u$ and $\boldsymbol{g}_u$ using a Taylor expansion of e^t up to a chosen order d. Replacing e^t by $\sum_{\ell=0}^{d} t^\ell/\ell!$ in Equations (4.10) and (4.12) and keeping notation $\boldsymbol{\zeta}_u$ and $\boldsymbol{g}_u$ for simplicity, Equations (4.9) and (4.11) become respectively

$$\sum_{\ell=0}^{d} \mathbb{E}[N^\ell]\frac{t^\ell}{\ell!} = \left[\sum_X \prod_{u=1}^{n} \boldsymbol{\zeta}_u(X_{\{\mathrm{pa}(u),u\}})\right]\Big|_d = \sum_X \overset{n}{\underset{u=1}{\bigstar}} \boldsymbol{\zeta}_u(X_{\{\mathrm{pa}(u),u\}}) \tag{4.13}$$

and

$$\sum_{\ell=0}^{d} \mathbb{E}[N^\ell|\mathrm{ev}]\frac{t^\ell}{\ell!} = \frac{1}{\mathbb{P}(\mathrm{ev})}\left[\sum_X \prod_{u=1}^{n} \boldsymbol{g}_u(X_{\{\mathrm{pa}(u),u\}})\right]\Big|_d = \frac{1}{\mathbb{P}(\mathrm{ev})}\sum_X \overset{n}{\underset{u=1}{\bigstar}} \boldsymbol{g}_u(X_{\{\mathrm{pa}(u),u\}}) \tag{4.14}$$

where $\cdot|_d$ denotes the truncation of a polynomial at the order d, $\bigstar$ is the conventional product with an additional truncation of the resulting polynomial at the order d and $\mathbb{P}(\mathrm{ev})$ is given by the first coefficient in $\sum_X \bigstar_{u=1}^{n} \boldsymbol{g}_u(X_{\{\mathrm{pa}(u),u\}})$. With the same reasoning Equations (4.13) and (4.14) can be computed with an upward pass of the sum-product algorithm in $\mathcal{O}\left(d \times m \times \max_u |\mathcal{X}_u|^{\tau+1}\right)$ time complexity.

Remark. The noticeable advantage of this method over an alternative one which consists in computing $\mathbb{E}\left[N^k|\mathrm{ev}\right]$ using $\mathrm{pgf}_{N|\mathrm{ev}}$ is its lower computational cost: $\mathcal{O}(c\times d)$ where $c = m \times \max_u |\mathcal{X}_u|^{\tau+1}$ for $\mathrm{mgf}_{N|\mathrm{ev}}$ truncated at a chosen order d versus $\mathcal{O}(c \times K)$ where K is the maximal value taken by N for $\mathrm{pgf}_{N|\mathrm{ev}}$.

4.4 Illustration

In this section, we briefly illustrate the method with a simulated Markov chain with errors. Let us consider an HMM composed of a Markov chain of hidden variables $X = (X_1, \dots, X_n) \in \{0,1\}^n$ and a set of observed variables $O = (O_1, \dots, O_n) \in \{0,1\}^n$ such that, for all $i \in \{1, \dots, n\}$, $\mathbb{P}(O_i \neq X_i | X_i) = \eta$ where η is a fixed parameter of error. We consider the evidence $\{O = \omega\}$ where $\omega = (\omega_1, \dots, \omega_n)$ is the vector of observed values taken by O and we denote by $N = \sum_{i=1}^n \mathbb{1}_{\{X_i \neq \omega_i\}}$ the number of errors. Let us introduce two sets of potentials $\{\boldsymbol{\phi}_i^p\}_{i=1,\dots,n}$ and $\{\boldsymbol{\phi}_i^m\}_{i=1,\dots,n}$ defined as

$$\boldsymbol{\phi}_1^p(X_1) = \mathbb{P}(X_1)(1-\eta)^{\mathbb{1}_{\{X_1=\omega_1\}}}(\eta z)^{\mathbb{1}_{\{X_1\neq\omega_1\}}},$$

$$\boldsymbol{\phi}_1^m(X_1) = \mathbb{P}(X_1)(1-\eta)^{\mathbb{1}_{\{X_1=\omega_1\}}}\left(\eta\sum_{\ell=0}^{d}\frac{t^\ell}{\ell!}\right)^{\mathbb{1}_{\{X_1\neq\omega_1\}}}$$

and for $i = 2, \dots, n$,

$$\boldsymbol{\phi}_i^p(X_{i-1}, X_i) = \mathbb{P}(X_i|X_{i-1})(1-\eta)^{\mathbb{1}_{\{X_i=\omega_i\}}}(\eta z)^{\mathbb{1}_{\{X_i\neq\omega_i\}}}$$

$$\boldsymbol{\phi}_i^m(X_{i-1}, X_i) = \mathbb{P}(X_i|X_{i-1})(1-\eta)^{\mathbb{1}_{\{X_i=\omega_i\}}}\left(\eta\sum_{\ell=0}^{d}\frac{t^\ell}{\ell!}\right)^{\mathbb{1}_{\{X_i\neq\omega_i\}}}$$

where z and t are two dummy variables and d is a chosen maximal order moment. Let $k_{\max}$ be an arbitrarily chosen maximal number of errors, note that we have

$$\sum_{k=0}^{k_{\max}} \mathbb{P}(N = k, O = \omega) z^k = \sum_X \boldsymbol{\phi}_1^p(X_1) \overset{n}{\underset{i=2}{\stackrel{\star}{\star}}} \boldsymbol{\phi}_i^p(X_{i-1}, X_i) \tag{4.15}$$

and

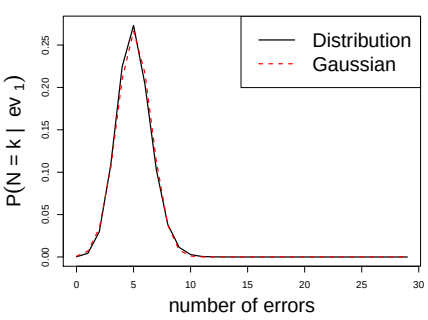
$$\sum_{\ell=0}^{d} \mathbb{E}\big[N^\ell, O = \omega\big]\frac{t^\ell}{\ell!} = \sum_X \boldsymbol{\phi}_1^m(X_1) \overset{n}{\underset{i=2}{\bigstar}} \boldsymbol{\phi}_i^m(X_{i-1}, X_i) \tag{4.16}$$

where $\stackrel{\star}{\star}$ and $\bigstar$ are the conventional product with an additional truncation step respectively at the order $k_{\max}$ and d.

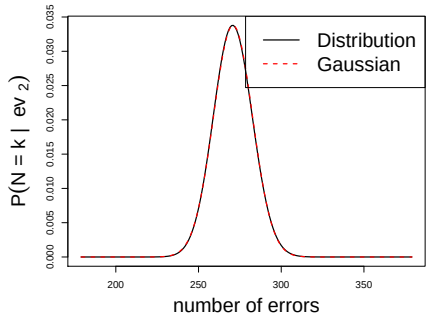
Equations (4.15) and (4.16) can be computed with an inward pass of the sum-product algorithm with initialization $\boldsymbol{F}_1^p(X_1) = \boldsymbol{\phi}_1^p(X_1)$ and $\boldsymbol{F}_1^m(X_1) = \boldsymbol{\phi}_1^m(X_1)$ and for $i = 2, \dots, n$, $\boldsymbol{F}_i^p(X_i) = \sum_{X_{i-1}} \boldsymbol{F}_{i-1}^p(X_{i-1}) \stackrel{\star}{\star} \boldsymbol{\phi}_i^p(X_{i-1}, X_i)$ and $\boldsymbol{F}_i^m(X_i) = \sum_{X_{i-1}} \boldsymbol{F}_{i-1}^m(X_{i-1}) \bigstar \boldsymbol{\phi}_i^m(X_{i-1}, X_i)$. The time complexity for each recursion is respectively of order $\mathcal{O}(n \times k_{\max})$ for Equation (4.15) and in $\mathcal{O}(n \times d)$ for Equation (4.16). At the end of each recursion we respectively have

$$\sum_{k=0}^{k_{\max}} \mathbb{P}(N = k, O = \omega) z^k = \sum_{X_n} \boldsymbol{F}_n^p(X_n) \quad \text{and} \quad \sum_{\ell=0}^{d} \mathbb{E}\big[N^\ell, O = \omega\big]\frac{t^\ell}{\ell!} = \sum_{X_n} \boldsymbol{F}_n^m(X_n).$$

Figure 4.1 displays results computed over two simulated sequences of respective length $n_1 = 50$ and $n_2 = 5000$ with parameters $\eta = 0.05$, $\mathbb{P}(X_1 = 0) = \mathbb{P}(X_1 =$



(a) $\mathbb{P}(N = k|\text{ev}_1)$ for the sequence of length $n_1 = 50$. $m_{1,1} = 5.04$ (dashed line), $m_{1,2} = 27.63$, $m_{1,3} = 162.61$, $m_{1,4} = 1018.81$.



(b) $\mathbb{P}(N = k|\text{ev}_2)$ for the sequence of length $n_2 = 5000$. $m_{2,1} = 270.85$ (dashed line), $m_{2,2} = 735.01 \times 10^2$, $m_{2,3} = 199.84 \times 10^5$, $m_{2,4} = 544.35 \times 10^7$.

Figure 4.1: Distribution of the number of errors (black line) and moments up to order 4 conditional the evidence ev_1 (left) and ev_2 (right) computed over a sequence of length $n_1 = 50$ letters (left) and $n_2 = 5000$ letters (right). For $j \in \{1, 2\}$, $m_{j,i}$ denotes $\mathbb{E}[N^i|\text{ev}_j]$. Additionally, the distribution of a Gaussian variable with mean $m_{j,1}$ and standard deviation $\sqrt{(m_{j,2} - m_{j,1}^2)}$ is drawn in red dashed line.

1) = 0.5, and for $i = 1, \ldots, 50$ (respectively $i = 1, \ldots, 5000$), for $r, s \in \{0, 1\}$, $\mathbb{P}(X_i = s|X_{i-1} = r) = \tau_{r,s}$ with $\tau = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}$. We denote by ev_j, the evidence for the sequence of length n_j, i.e. the set variables $(O_i)_{i=1,\ldots,n_j}$ and instantiations. A graphical representation of the distribution of the number of errors conditional on ev_j is proposed along with the density of a Gaussian variable with mean and standard deviation obtained from computed moments. The i-th order moments are also given up to $i = 4$ and were verified with the alternative method mentioned in the remark at the end of Section 4.3.

Chapter 5

Exact derivatives of the likelihood in Bayesian networks with polynomial potentials

Sommaire

5.1 Introduction

This chapter constitute the first contribution of the thesis. Following ideas based on generating functions in Bayesian networks (BNs) (see Section 4.1), we developed a method for computing the exact derivatives of the likelihood in a parametric BN up to a chosen order d in $\mathcal{O}(c\times d^2)$ in time for a unidimensional parameter and $\mathcal{O}(c\times p^{2d})$ in time for a p-dimensional parameter, $p>1$, where c is the complexity for computing the likelihood itself. These complexities are similar to numerical methods with the main advantage that we obtain exact derivatives instead of approximations.

We consider in this section a BN $\mathcal{B}=(G=(X=\{X_1,\dots,X_n\},\mathcal{E}),\mathbb{P})$ parametrized by $\theta\in\mathbb{R}^p$. For all $u\in\{1,\dots,n\}$, we denote by $\mathcal{X}_u$ the state space of X_u. Let $\text{ev}\overset{\text{def}}{=}\cap_{u\in E}\{X_u\in\mathcal{X}_u^*\subset\mathcal{X}_u\}$ where $E\subset\{1,\dots,n\}$ be an evidence for $\mathcal{B}$ and, for all $u\in\{1,\dots,n\}\setminus E$, let $\mathcal{X}_u^*=\mathcal{X}_u$, the likelihood of θ is given by

$$L(\theta)=\mathbb{P}(\text{ev}|\theta)=\sum_{X_1}\dots\sum_{X_n}\prod_{u=1}^{n}\phi_u\left(X_{\{\text{pa}(u),u\}}|\theta\right) \tag{5.1}$$

where $\text{pa}(u)$ is the set of indexes of graph parents of X_u in G, for $U \in \{1, \ldots, n\}$, $X_U = \{X_u\}_{u \in U}$ and $\phi_u(X_{\{\text{pa}(u),u\}}|\theta) = \mathbb{1}_{\{X_v \in \mathcal{X}_v^*, \forall v \in \{\text{pa}(u),u\}}\mathbb{P}(X_u|X_{\text{pa}(u)})$. Computing the derivatives of the log-likelihood function is of great interest, especially the first and second order derivatives, from which one can derive the score and the observed Fisher information matrix respectively defined as the first order derivative and the negative of the second order derivative of the log-likelihood. These quantities can not only help maximizing the likelihood function (e.g. through Newton-based algorithms) but can also be used for several other tasks of interest such as the computation of confidence intervals for parameters as well as performing hypothesis testing (Prum, 2010).

In sensitivity analysis, one questions the sensitivity of a query to small variations in certain network parameters (see Jensen and Nielsen, 2007, pp 184–185). Much focus has been put on developing efficient algorithms (Castillo et al., 1997; Chan and Darwiche, 2002, 2012). In such analysis a parameter is an entry in a local probability distribution of the form $\mathbb{P}(X_u|X_{\text{pa}(u)})$. A potential is expressed as a polynomial in θ. However when the same parameter appears in many potentials, the resulting polynomial is usually of high order, and its computational cost prohibitive. As an extension to sensitivity analysis Darwiche (2003) uses a network polynomial with network parameters being the potentials. The authors introduce a multilinear function containing two types of variables (evidence indicators and network parameters) and apply arithmetic circuits for an efficient computation. However such method can only deal with a simple parametrization (e.g. one parameter for each probability table entry).

Due to the importance of second order derivatives of the log-likelihood, some authors proposed methods for calculating it, in particular when the Expectation-Maximization (EM) algorithm is used for optimizing the log-likelihood. For instance Oakes (1999) states that

$$\frac{\partial \log L(\theta)}{\partial \theta} = \left\{ \frac{\partial Q(\theta'|\theta)}{\partial \theta'} \right\}_{\theta'=\theta}$$

where $Q(\theta'|\theta) = \sum_X \mathbb{P}(X|\text{ev};\theta) \log \mathbb{P}(X, \text{ev}|\theta')$ is the auxiliary function of the EM algorithm given by $Q(\theta'|\theta) = \sum_{u=1}^n \sum_{X_{\text{pa}(u),u}} \mathbb{P}(X_{\{\text{pa}(u),u\}}|\text{ev};\theta) \log \phi_u(X_{\text{pa}(u)}, X_u|\theta')$ in our context. Therefore we have

$$\frac{\partial \log L(\theta)}{\partial \theta} = \sum_{u=1}^{n} \sum_{X_{\{\text{pa}(u),u\}}} \mathbb{P}(X_{\{\text{pa}(u),u\}}|\text{ev};\theta) \frac{\partial \log \phi_u(X_{\{\text{pa}(u),u\}}|\theta')}{\partial \theta} \tag{5.2}$$

and

$$\frac{\partial^2 \log L(\theta)}{\partial \theta^2} = \sum_{u=1}^{n} \sum_{X_{\{\text{pa}(u),u\}}} \left\{ \frac{\partial \mathbb{P}(X_{\{\text{pa}(u),u\}}|\text{ev};\theta)}{\partial \theta} \frac{\partial \log \phi_u(X_{\{\text{pa}(u),u\}}|\theta)}{\partial \theta^T} + P(X_{\{\text{pa}(u),u\}}|\text{ev};\theta) \frac{\partial^2 \log \phi_u(X_{\{\text{pa}(u),u\}}|\theta)}{\partial \theta^2} \right\} \tag{5.3}$$

Each quantity in Equations (5.2) and (5.3) are either straightforwardly implemented or obtained via an inward pass of the sum-product algorithm in linear in c time

complexity except $\left[\partial \mathbb{P}(X_{\{\mathrm{pa}(u),u\}}|\mathrm{ev};\theta)/\partial\theta\right]$. Application examples selected in Oakes (1999) are restricted to exponential families and a simple genetic example for which we get closed *formulae*.

Another approach was developed by Louis (1982) who proved that the second order derivative of the log-likelihood is given by the following sum of conditional expectation of additive functionals named as the Louis' identities. The authors state that

$$\frac{\partial^2 \log L(\theta)}{\partial\theta^2} = -\frac{\partial \log L(\theta)}{\partial\theta}\frac{\partial \log L(\theta)}{\partial\theta^T} + \mathbb{E}\left[\frac{\partial^2 \log \mathbb{P}(X,\mathrm{ev}|\theta)}{\partial\theta^2}\middle|\mathrm{ev};\theta\right] + \mathbb{E}\left[\frac{\partial \log \mathbb{P}(X,\mathrm{ev}|\theta)}{\partial\theta}\frac{\partial \log \mathbb{P}(X,\mathrm{ev}|\theta)}{\partial\theta^T}\middle|\ \mathrm{ev};\theta\right] \quad (5.4)$$

where $\partial \log \mathbb{P}(X,\mathrm{ev}|\theta)/\partial\theta = \sum_{u=1}^{n} \partial\theta \log \phi_u(X_{\{\mathrm{pa}(u),u\}}|\theta)/\partial\theta$, $\partial^2 \log \mathbb{P}(X,\mathrm{ev}|\theta)/\partial\theta^2 = \sum_{u=1}^{n} \partial^2 \log \phi_u(X_{\{\mathrm{pa}(u),u\}}|\theta)/\partial\theta^2$ and exponent T stands for transpose. The last line of Equation (5.4) involves the sum of a square functional for which Cappé and Moulines (2005) offer a method based on *smoothing recursions* in the particular framework of Hidden Markov Models (HMMs). Alternatively, one could use the method detailed in (Cowell, 1992; Nilsson, 2001; Nuel, 2010) to compute the first and second order moments of these additive functionals.

In this chapter, we would like to propose a novel approach based on generating functions to compute the derivatives of the likelihood in a BN up to a chosen order. Our method can be viewed as an extension of the work of Cowell (1992) and Nilsson (2001) to derivatives.

This chapter is organized as follows: in Section 5.2, we start by defining functions and operators needed to implement our method before explaining in details its implementation. In Section 5.3, we propose two illustrative application examples. The first one is a simple toy binary BN. The second one is taken from two-point linkage models used in genetic analysis for localizing a targeted gene on the genome. Finally, in Section 5.4, we discuss possible extensions, comparisons and combinations with existing methods as main perspectives.

This work was published in *Proceedings of Machine Learning Research* (Lefebvre and Nuel, 2018) with a notation slightly different from the one used in Chapter 1. Several modifications where performed in order to better suit Chapter 1, however some conflicting points and redundancies remain unavoidable, in particular, sets of indexes are introduced in the present chapter whereas Chapter 1 sticks to sets of variables.

5.2 Method

Before detailing the implementation of the method, we first need to introduce two new definitions.

Definition 10 (derivative generating function)**.** *Let f be a function of class $\mathcal{C}^d$ ($d \in \mathbb{N}$) of $\theta \in \mathbb{R}$, we define the derivative generating function of f as the generating*

function associated with the sequence of its derivatives:

$$D^d f(\theta) = \sum_{k=0}^{d} f^{(k)}(\theta) z^k$$

where z is a dummy variable.

Remark: One can generalize the derivative generating function to a multidimensional parameter with the sequence of partial derivatives. Let f be a function of class $\mathcal{C}^d$ of $\theta = \{\theta_1, \ldots, \theta_p\} \in \mathbb{R}^p$, we define the derivative generating function of f to be

$$D^d f(\theta) = \sum_{\substack{k_1,\ldots,k_p \\ k_1+\ldots+k_p \leq d}} \frac{\partial^{(k_1+\ldots+k_p)} f(\theta)}{\partial {\theta_1}^{k_1} \ldots \partial {\theta_p}^{k_p}} {z_1}^{k_1} \ldots {z_p}^{k_p}$$

where $z_1, \ldots, z_p$ are p dummy variables.

Our aim is to apply the sum-product algorithm over polynomial potentials to compute $D^d L(\theta) = \sum_{k=0}^{d} L^{(k)}(\theta) z^k$ up to an arbitrary order d. For the sake of simplicity, we will focus on the unidimensional case and briefly extend the notions to a multidimensional parameter at the end of the section.

We need to adapt the multiplicative law and for that extend, let us introduce the following new definition.

Definition 11 (Leibniz product)**.** *Let $P = \sum_{k=0}^{d} a_k z^k$ and $Q = \sum_{k=0}^{d} b_k z^k$ be two polynomials in z, we define the Leibniz product of P and Q as*

$$P \star Q = \sum_{k=0}^{d} \sum_{i=0}^{k} \binom{k}{i} a_{k-i} b_i z^k$$

where $\binom{k}{i}$ is the binomial coefficient. Note that we deliberately drop all coefficients of degree greater than d.

From these two definitions, we can straightforwardly derive the following proposition.

Proposition 4. *Let f and g be two functions of class $\mathcal{C}^d$ of $\theta \in \mathbb{R}$,*

$$D^d f(\theta) \star D^d g(\theta) = D^d (fg)(\theta).$$

Proof. Let f and g be two functions of class $\mathcal{C}^d$ of $\theta \in \mathbb{R}$. Let $P = D^d f(\theta) = \sum_{k=0}^{d} f^{(k)}(\theta) z^k$ and $Q = D^d g(\theta) = \sum_{k=0}^{d} g^{(k)}(\theta) z^k$, then

$$P \star Q = \sum_{k=0}^{d} \sum_{i=0}^{k} \binom{k}{i} f^{(k-i)}(\theta) g^{(i)}(\theta) z^k.$$

We recognize the general Leibniz rule for computing the derivatives of the product of two functions which concludes the proof. □

Let us now highlight how the sum-product algorithm can be applied to reduce the complexity to compute $D^dL(\theta) = \sum_{k=0}^{d} L^{(k)}(\theta)z^k$ up to a chosen order d. Returning to Chapter 1 and in particular to Proposition 3, let $\phi = \{D^d\phi_u\}_{u=1,\dots,n}$ be the set of derivative generating functions up to the order d of potentials ϕ_u introduced in Equation (5.1), note that, for all $u \in \{1,\dots,n\}$, $D^d(\phi_u)$ is a potential and $\text{Scope}(D^d\phi_u) = \text{Scope}(\phi_u) = X_{\{\text{pa}(u),u\}}$. We denote by $J = ((C_1,\dots,C_m),\mathcal{F})$, a junction-tree defined by $\text{VE}(\phi, X, \sigma)$ where σ is an elimination ordering over X (see Proposition 3). Returning to Section 1.4.1, let C_m be the chosen root clique in J, for all $i \in 1,\dots m$, we denote by $\mathcal{C}_i \subset \{1,\dots,n\}$ (respectively $\mathcal{S}_i \subset \{1,\dots,n\}$) the set of labels of the variables in the i-th clique (respectively i-th separator) of J. Therefore we have $C_i = \{X_u, u \in \mathcal{C}_i\}$ and $S_i = \{X_u, u \in \mathcal{S}_i\}$. Let of_u be the choice of a unique $i \in \{1,\dots,m\}$ such that $X_{\{\text{pa}(u),u\}} \subset C_i$, we say that $D^d\phi_u$ is injected in $C_{\text{of}(u)}$ and we introduce $\mathcal{C}_i{}^* = \{u \in \{1,\dots,n\}, \text{of}_u = i\}$. Then, adapting Definition 1.9, the polynomial potentials of each clique C_i for an arbitrary order d is defined as:

$$\mathbf{\Phi}_i^d(C_i|\theta) = \bigstar_{u\in\mathcal{C}_i{}^*} D^d\phi_u(X_{\{\text{pa}(u),u\}}|\theta). \tag{5.5}$$

For all $i \in \{1,\dots,m\}$, let $\text{to}_i \in \{i+1,\dots,m\}$ be the label of the subsequent clique of C_i. For all $i \in \{1,\dots,m\}$, we define V_i to be the set of cliques upstream C_i, C_i included, and $\mathcal{V}_i = \{j, C_j \in V_i\}$. Then, for all $i \in \{1,\dots,m\}$, we define the polynomial *inward messages* to be

$$\boldsymbol{\delta}_i^d(S_i|\theta) = \sum_{V_i\setminus S_i} \bigstar_{u\in\mathcal{V}_i{}^*} D^d\phi_u(X_{\{\text{pa}(u),u\}}|\theta)$$

where $\mathcal{V}_i{}^* = \{u \in \{1,\dots,n\}, \exists j \in \mathcal{V}_i, \text{of}_u = j\}$. Note that $V_m = X_{\mathcal{V}_m^*} = X$ and $S_m = \emptyset$ and we get in particular:

$$\boldsymbol{\delta}_m^d(\emptyset|\theta) = \sum_X \bigstar_{u=1}^{n} D^d\phi_u(X_{\{\text{pa}(u),u\}}|\theta) = D^d\left(\sum_X \prod_{u=1}^{n} \phi_u(X_{\{\text{pa}(u),u\}}|\theta)\right) = D^dL(\theta).$$

One can apply the message passing algorithm over polynomial potentials to recursively compute the *inward messages* using the following proposition:

Proposition 5. $\forall i \in \{1,\dots m\}$,

$$\boldsymbol{\delta}_i^d(S_i|\theta) = \sum_{C_i\setminus S_i} \left(\bigstar_{j\in\text{nb}_i\setminus\text{to}_i} \boldsymbol{\delta}_j^d(S_j|\theta)\right) \star \mathbf{\Phi}_i^d(C_i|\theta).$$

where nb_i *is the set of indexes of neighbors of* C_i.

Proof. The proof is straightforward when considering the *belief propagation* in J where the additive law is the conventional additive law ($+$) and the multiplicative law is the Leibniz product ($\bigstar$). Some details of the proof are given below: For all $i \in \{1,\dots m\}$,

$$\boldsymbol{\delta}_i^d(S_i|\theta) = \sum_{V_i\setminus S_i} \bigstar_{u\in\mathcal{V}_i{}^*} D^d\phi_u(X_{\{\text{pa}(u),u\}}|\theta) = \sum_{\substack{V_j\setminus S_j \\ j\in\text{nb}(i)\setminus\text{to}(i)}} \sum_{C_i\setminus S_i} \bigstar_{u\in\mathcal{V}_i{}^*} D^d\phi_u(X_{\{\text{pa}(u),u\}}|\theta).$$

Recalling the properties of a junction-tree we have, for all $i \in \{1,\dots,m\}, V_i \setminus S_i = \sqcup_{j\in \mathrm{nb}_i\setminus \mathrm{to}(i)} V_j \setminus S_j \sqcup C_i \setminus S_i$ and $\mathcal{V}_i{}^* = \sqcup_{j\in \mathrm{nb}_i\setminus \mathrm{to}(i)} \mathcal{V}_j{}^* \sqcup \mathcal{C}_i{}^*$ where $\sqcup$ is the disjoint union, and therefore

$$\boldsymbol{\delta}_i^d(S_i|\theta) = \sum_{C_i\setminus S_i} \left(\bigstar_{j\in \mathrm{nb}(i)\setminus \mathrm{to}(i)} \underbrace{\sum_{V_j\setminus S_j} \left(\bigstar_{u\in \mathcal{V}_j{}^*} D^d\phi_u(X_{\{\mathrm{pa}(u),u\}}|\theta) \right)}_{\boldsymbol{\delta}_j^d(S_j|\theta)} \right) \bigstar_{u\in \mathcal{C}_i{}^*} D^d\phi_u(X_{\{\mathrm{pa}(u),u\}}|\theta)$$

which concludes the proof by induction.

□

The recursive computation of inward messages with the sum-product algorithm is of the order of $\mathcal{O}(c \times d^2)$ in time where c is the complexity of an inference over a JT defined by $\mathrm{VE}(\{\phi_u\}_{u=1,\dots,n}, X, \sigma)$.

Remark: We make the choice here to focus on inward messages as we are only interested in the likelihood and its derivatives but one can add a backward recursion to compute marginal and joint probabilities and their derivatives.

The extension to a multidimensional parameter $\theta = \{\theta_1,\dots,\theta_p\} \in \mathbb{R}^p$, implies multivariate polynomials with as many dummy variables as dimensions of the parameter. The Leibniz product of two multivariate polynomials is defined as:

$$P \star Q = \sum_{\substack{k_1,\dots,k_p \\ k_1+\dots+k_p\le d}} \sum_{i_1=0}^{k_1} \dots \sum_{i_p=0}^{k_p} \binom{k_1}{i_1} \dots \binom{k_p}{i_p} a_{k_1-i_1,\dots,k_p-i_p} b_{i_1,\dots,i_p} {z_1}^{k_1} \dots {z_p}^{k_p}$$

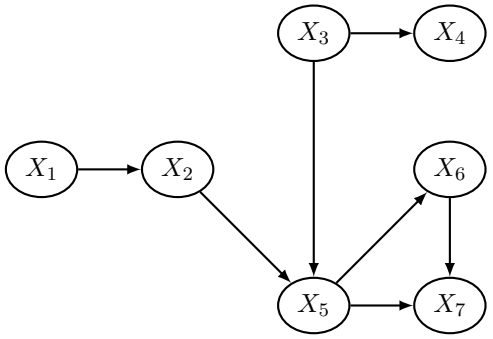
where f is a function of class $\mathcal{C}^d$ of $\theta \in \mathbb{R}^p$ and P and Q are two polynomials of degree at most d in p dummy variables. The generalization of Proposition 5 to a multidimensional parameter gives:

$$\boldsymbol{\delta}_m^d(\emptyset|\theta) = \sum_{\substack{k_1,\dots,k_p \\ k_1+\dots+k_p\le d}} \frac{\partial^{(k_1+\dots+k_p)} L(\theta)}{\partial {\theta_1}^{k_1} \dots \partial {\theta_p}^{k_p}} {z_1}^{k_1} \dots {z_p}^{k_p} = D^d L(\theta)$$

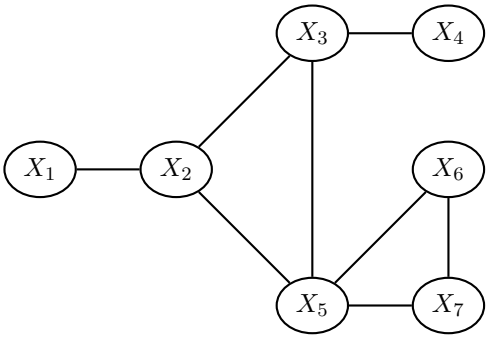
which is obtained in $\mathcal{O}(c \times p^{2d})$ in time.

5.3 Results

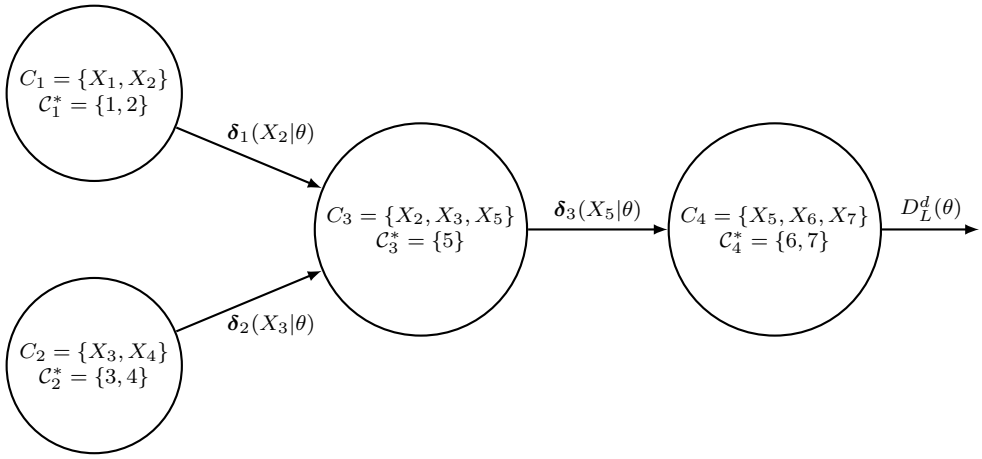
This illustrative section is composed of two application examples. The first one is a simple toy-example with a BN over binary variables and the second one is taken from two-point linkage analysis which aims at locating a targeted gene on the genome. For the sake of simplicity, we restrict this section to a unidimensional parameter and we consider the maximal order $d = 2$.



(a) Directed acyclic graph



(b) Moral graph



(c) Junction-tree J

Figure 5.1: Directed acyclic graph and its moral graph associated with the toy example (respectively top left and top right) and a junction-tree J defined by $\mathrm{VE}(\phi, X, \sigma)$ where $\phi = \{D^d\phi_u\}_{u=1,\dots,7}$ for any $d \geq 0$ and $\sigma = (X_1, X_4, \{X_2, X_3\}, X_5, X_6, X_7)$ is perfect for the moral graph (bottom). For all $i \in \{1, \dots, 4\}$, $\mathcal{C}_i^*$ denotes the set of indexes of potentials injected in clique C_i.

5.3.1 Toy-example: a Bayesian network over binary variables

Let us firstly consider the BN $\mathcal{B}$ over n variables $X = \{X_1, \dots, X_n\}$ with probability distribution $\mathbb{P}$ that factorizes over the Directed Acyclic Graph (DAG) represented in Figure 5.1a where $n = 7$ and $X \in \{0, 1\}^n$. Its moral graph is pictured in Figure 5.1b. For all $u \in \{1, \dots, n\}$, we assume that

$$\mathbb{P}(X_u = 1 | X_{\mathrm{pa}(u)}; \theta) = \frac{\exp\left(\mu + \theta \sum_{v \in \mathrm{pa}(u)} X_v\right)}{1 + \exp\left(\mu + \theta \sum_{v \in \mathrm{pa}(u)} X_v\right)},$$

thus

$$\mathbb{P}\left(X_u = 1 \middle| \sum_{v\in \mathrm{pa}(u)} X_v = k; \theta\right) = \frac{e^{(\mu+\theta k)}}{1+e^{(\mu+\theta k)}}$$

where $\mu = -0.5$ is assumed to be known. Let us define

$$P_k = f_k(\theta) + f_k'(\theta)z + f_k''(\theta)z^2 \quad \text{with} \quad f_k(\theta) = \frac{e^{(\mu+k\theta)}}{1+e^{(\mu+k\theta)}}$$

where a prime symbol denotes the derivative with respect to θ.

Let $\mathrm{ev} \overset{\text{def}}{=} \cap_{u\in E}\{X_u \in \mathcal{X}_u^* \subset \mathcal{X}_u\}$ where $E \subset \{1,\dots,n\}$ be an evidence for $\mathcal{B}$ and, for all $u \in \{1,\dots,n\} \setminus E$, let $\mathcal{X}_u^* = \mathcal{X}_u$, noticing that $1 - P_k = D^2(1 - f_k(\theta))$, the polynomial potential associated with $\mathbb{P}(X_u = 1|X_{\mathrm{pa}(u)};\theta)$ is given by the expression:

$$D^2\phi_u\left(X_{\{\mathrm{pa}(u),u\}}|\theta\right) = \begin{cases} \mathbb{1}_{\{X_v\in\mathcal{X}_v^*,\forall v\in\{\mathrm{pa}(u),u\}\}} \times P_{\sum_{v\in\mathrm{pa}(u)} X_v} & \text{if } X_u = 1 \\ \mathbb{1}_{\{X_v\in\mathcal{X}_v^*,\forall v\in\{\mathrm{pa}(u),u\}\}} \times \left(1 - P_{\sum_{v\in\mathrm{pa}(u)} X_v}\right) & \text{if } X_u = 0 \end{cases}$$

where ϕ_u is defined as in Equation (5.1). A graphical representation of a JT denoted J defined by $\mathrm{VE}(\{D^d\phi_u\}_{u=1,\dots,n}, X, \sigma)$ is given in Figure 5.1c where $\sigma = (X_1, X_4, \{X_2, X_3\}, \{X_5, X_6, X_7\})$ is perfect for the moral graph 5.1b. For each $i \in \{1,\dots,4\}$, the set $\mathcal{C}_i^*$ and message $\boldsymbol{\delta}_i$ associated with J is added on the graph.

We consider entering the evidence $\mathrm{ev} = \{X_1 = 0, X_7 = 1\}$ in the BN and choose maximal order $d = 2$. Thus we get $D^2\phi_1(X_1 = 1) = 0$ and, for all $x \in \mathcal{X}_5^*$ and $y \in \mathcal{X}_6^*$, $D^2\phi_7(X_5 = x, X_6 = y, X_7 = 0|\theta) = 0$, the null polynomial. For all $i \in \{1,\dots,4\}$, polynomial potentials $\boldsymbol{\Phi}_i^2(C_i|\theta)$ and inward messages $\boldsymbol{\delta}_i^2(S_i|\theta)$ are respectively computed with Equation (5.5) and recursively with Proposition 5. For example, dropping θ for a lighter notation and assuming that $\boldsymbol{\Phi}_1^2(X_1, X_2)$, $\boldsymbol{\Phi}_2^2(X_3, X_4)$, $\boldsymbol{\delta}_1^2(X_2)$ and $\boldsymbol{\delta}_2^2(X_3)$ are computed, quantities $\boldsymbol{\Phi}_3^2(X_2, X_3, X_5)$ and $\boldsymbol{\delta}_3^2(X_5)$ can respectively be written as

$$\boldsymbol{\Phi}_3^2(X_2, X_3, X_5) = D^2\phi_5(X_2, X_3, X_5)$$

and

$$\boldsymbol{\delta}_3^2(X_5) = \sum_{X_2}\sum_{X_3} \boldsymbol{\delta}_1^2(X_2) \star \boldsymbol{\delta}_2^2(X_3) \star \boldsymbol{\Phi}_3^2(X_2, X_3, X_5).$$

Table 5.1 gives the expression of a few chosen polynomial clique potentials and polynomial inward messages in J for $\theta = 1$ and $\mathrm{ev} = \{X_1 = 0, X_7 = 1\}$. Note that, because X_3 has no parent and the only parent of X_4 is X_3, potentials $\boldsymbol{\Phi}_2^2(X_3 = 0, X_4)$ for $X_4 \in \{0, 1\}$ are of degree 0. Moreover, because X_4 is not instantiated in ev and X_3 has no parent, $\boldsymbol{\delta}_2^2(X_3 = 0) + \boldsymbol{\delta}_2^2(X_3 = 1) = 1$. All other potentials in this table are of degree greater than 0. In particular $\boldsymbol{\delta}_4^2(\emptyset)$ which has been verified numerically (data not shown) is a polynomial in z containing $L(\theta)$ and its derivatives up to the order 2 in its coefficients.

The log-likelihood of θ, $\ell(\theta) = \log L(\theta)$, and its derivatives up to the order 2 evaluated at $\theta = 1$ are listed in Table 5.2 for different simulations of N values for $\{X_1, \dots, X_7\}$ using true parameter $\theta^* = 1$. We denote by ev_{ab} the observation $\{X_1 = a, X_7 = b\}$ and N_{ab} the number of times ev_{ab} is observed among the N

$\mathbf{\Phi}_2^2(X_3 = 0, X_4 = 0)$	0.3874556
$\mathbf{\Phi}_2^2(X_3 = 0, X_4 = 1)$	0.2350037
$\mathbf{\Phi}_2^2(X_3 = 1, X_4 = 0)$	$0.142537 - 0.08872346z + 0.02173003z^2$
$\mathbf{\Phi}_2^2(X_3 = 1, X_4 = 1)$	$0.2350037 + 0.08872346z - 0.02173003z^2$
$\boldsymbol{\delta}_2^2(X_3 = 0)$	0.6224593
$\boldsymbol{\delta}_2^2(X_3 = 1)$	0.3775407
$\boldsymbol{\delta}_3^2(X_5 = 0)$	$0.2767608 - 0.09521838z + 0.05045801z^2$
$\boldsymbol{\delta}_3^2(X_5 = 1)$	$0.3456986 + 0.09521838z - 0.05045801z^2$
$\boldsymbol{\delta}_4^2(\emptyset) = D^2L(\theta)$	$0.3872484 + 0.1613463z - 0.05839165z^2$

Table 5.1: A sample of chosen clique potentials and inward messages in J.

simulations. As expected, because each observation contributes independently to the likelihood, $\ell(1)$ decreases linearly with N. Moreover, regarding the negative of the second order derivative of $\ell(\theta)$ evaluated at $\theta = 1$, note that $-\ell''(1)$ is low for N=1 leading to a large variance which was expected in the framework of a single observation of the couple $\{X_1, X_7\}$. Furthermore $-\ell''(1)$ increases linearly with an increasing N as expected as the Cramer-Rao bound for the variance of θ defined as $1/(-\ell''(\theta))$ decreases linearly with N.

$\boldsymbol{N}$	$\boldsymbol{N_{00}}$	$\boldsymbol{N_{01}}$	$\boldsymbol{N_{10}}$	$\boldsymbol{N_{11}}$	$\boldsymbol{\ell(1)}$	$\boldsymbol{\ell'(1)}$	$\boldsymbol{\ell''(1)}$
1	1	0	0	0	-1.447	2.483×10^{-1}	-7.476×10^{-1}
1	0	1	0	0	-9.487×10^{-1}	-1.508×10^{-1}	3.939×10^{-1}
1	0	0	1	0	-1.988	1.493×10^{-1}	-8.439×10^{-1}
1	0	0	0	1	-1.425	-0.850×10^{-1}	4.604×10^{-1}
50	4	20	9	17	-6.688×10^{1}	6.144	-1.727×10^{1}
500	134	189	61	116	-6.598×10^{2}	-0.905×10^{1}	-1.584×10^{2}
5000	1212	1854	767	1167	-6.700×10^{3}	-1.434×10^{2}	-1.628×10^{3}
50,000	11770	19292	6844	12094	-6.617×10^{4}	-3.722	-1.614×10^{4}

Table 5.2: Log-likelihood and its derivatives up to the order 2 computed for different simulations of N values for $\{X_1, \ldots, X_7\}$ leading to N_{ab} observed couples $\{X_1 = a, X_7 = b\}$.

5.3.2 Two-point linkage in genetics

We propose in this section an application in two-point linkage analysis, a field in statistical genetics and genetic epidemiology which aims at locating a targeted gene on the genome by estimating the genetic distance between that gene and a maker of known localization (see Section 3.1 and in particular Section 3.1.3.2, paragraph "Linkage analysis", for recalls in genetics and genetic linkage analysis as well as Section 3.2 for the application of *belief propagation* in pedigrees). We also recommend the reference (Lauritzen and Sheehan, 2003, Section 2.2) for detailed explanations of basics in genetics and genetic linkage. Multi-point linkage analysis uses multiple

markers and usually leads to increased power and therefore it favored an extensive interest in statistical genetics. However we made the choice, in the present illustrative section, to focus on a single marker for easing results interpretation.

We consider the `contraol2a` example offered in the Mendel package (Lange et al., 2013) using PGM1 as the marker and RADIN as the targeted gene. The gene PGM1 has 4 alleles ($\{$`1+`, `1-`, `2+`, `2-`$\}$) with given allele frequencies as well as sensitivity and specificity of genetic testing, i.e. the probability of an observed sequenced marker conditional on the genotype. The gene RADIN is biallelic ($\{$`+`, `-`$\}$) with given allele frequencies and penetrance. We start this section with brief recalls of main notions in two-point linkage before expressing polynomial potentials in that context and propose a selection of computed results over the KUS family (N = 22 individuals) and the whole controle2a (N = 93 individuals) dataset of pedigrees.

Introduction the two-point linkage analysis. Most human cells are diploid which means that they contain pairs of chromosomes, one of paternal, one of maternal origin. Two chromosomes of the same pair are said to be homologous. In gonads, a diploid cell with double-stranded chromosomes splits into four haploid cells with single-stranded chromosomes called gametes, dedicated to be transmitted to an offspring. Homologous chromosomes can exchange genetic material during a meiosis and produce recombinant gametes. This phenomenon is called a crossover. A graphical representation of a crossover is given in Figure 5.2.

The closest two genes are on the genome, the less chances their alleles are separated during meiosis. Based on this phenomenon, a two-point linkage analysis uses results of genetic sequencing for a marker whose location is known on the genome, the penetrance of the targeted gene (probability of the trait (or phenotype) conditional on the genotype) and allele frequencies previously estimated with segregation analysis, in order to estimate the distance between the targeted gene denoted X and the marker denoted M. That distance is expressed as a function of the fraction of recombinant gametes $\theta = \#R/(\#R + \#NR)$. Variables involved in two-point linkage and their probabilistic relationships are pictured in the DAG represented in Figure 5.3 where, for $G \in \{X, M\}$, G^p (respectively G^m) stands for the paternal (respectively the maternal) allele for the gene G, for $h \in \{p, m\}$, SG^h denotes the selector (i.e. the origin) of G^h, Y denotes the phenotype coded by X and T, the genetic test for M. Each variable is indexed with its associated individual (1 for the father, 2 for the mother and 3 for the offspring). Selectors are binary variables and take their values in $\{p, m\}$ according to the origine of the associated allele. For instance $\mathrm{SX}_3^p = p$ (respectively $\mathrm{SX}_3^p = m$) if SX_3^p comes from the paternal (respectively maternal) chromosome of the father of Individual 3.

Polynomial potentials. For the sake of simplicity we will expose our method over a simple trio composed of one father, one mother and one child and we focus again on derivatives up to degree $d = 2$. Let $W = \{W_u\}_{u=1,\dots n}$ be the set of variables in Figure 5.3, we first need to implement polynomial potentials $D^2\mathbb{P}(W_u|W_{\mathrm{pa}_u}; \theta)$ for all $u \in \{1, \dots, n\}$. Note that, for all $W_u \in W \setminus \{\mathrm{SM}_3^p, \mathrm{SM}_3^m\}$, $D^2\mathbb{P}(W_u|W_{\mathrm{pa}_u}; \theta) = \mathbb{P}(W_u|W_{\mathrm{pa}_u})$ are all zero degree polynomials.

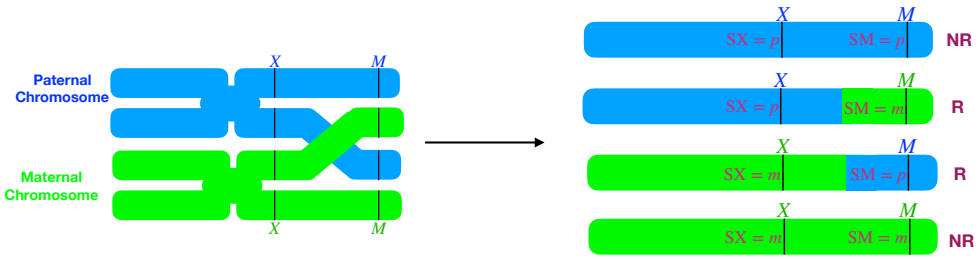


Figure 5.2: A crossover between two homologous chromosomes during a meiosis where SX (respectively SM) denotes the origine of X (respectively M). R (respectively NR) stands for recombinant (respectively non-recombinant) gamete.

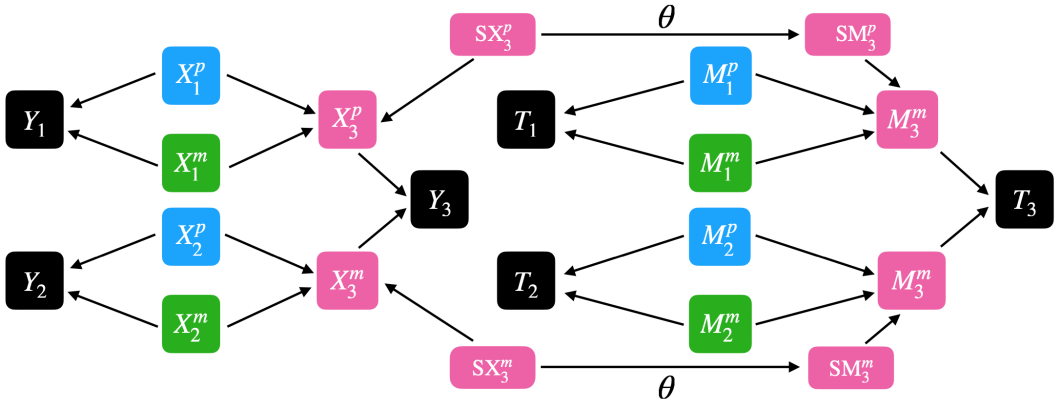


Figure 5.3: DAG of the variables involved in genetic linkage analysis for a simple trio composed of one father labeled 1, one mother labeled 2 and one child labeled 3. Y stands for the trait and T for the genetic test for the marker. The parameter θ appears in Conditional Probability Distributions (CPDs) of the form $\mathbb{P}(\mathrm{SM}_3^p|\mathrm{SX}_3^p;\theta)$ and $\mathbb{P}(\mathrm{SM}_3^m|\mathrm{SX}_3^m;\theta)$ where SX_3^p, SX_3^m, SM_3^p and SM_3^m are selectors for the associated allele.

We make two common assumptions in genetic analysis. Firstly, we assume that alleles of a common gene segregate independently such that, for $s \in \{p, m\}$, $D^d\mathbb{P}(\mathrm{SX}_3^p = s) = D^d\mathbb{P}(\mathrm{SX}_3^m = s) = 0.5$. Secondly we assume the genotypes of founders (individuals with no reported ancestors) are in Hardy-Weinberg equilibrium (constant allele frequencies from a generation to the next one). Under above assumptions, for all $W_u \in W \setminus \{\mathrm{SX}_3^p, \mathrm{SX}_3^m, \mathrm{SM}_3^p, \mathrm{SM}_3^m\}$, the implementation of $D^d\mathbb{P}(W_u|W_{\mathrm{pa}_u}; \theta) = \mathbb{P}(W_u|W_{\mathrm{pa}_u})$ is straightforward using allele frequencies, sensitivity and specificity of genetic tests and penetrance of the targeted gene.

Both potentials $\mathbb{P}(\mathrm{SM}_3^p|\mathrm{SX}_3^p; \theta)$ and $\mathbb{P}(\mathrm{SM}_3^m|\mathrm{SX}_3^m; \theta)$ depend on θ and therefore, their derivative generating function up to degree $d = 2$ is a polynomial of strictly positive degree and we have

$$D^2\mathbb{P}(\mathrm{SM}_3^p = \mathrm{SX}_3^p|\mathrm{SX}_3^p; \theta) = D^2\mathbb{P}(\mathrm{SM}_3^m = \mathrm{SX}_3^m|\mathrm{SX}_3^m; \theta) = (1 - \theta) - z$$

and

$$D^2\mathbb{P}(\mathrm{SM}_3^p \neq \mathrm{SX}_3^p|\mathrm{SX}_3^p; \theta) = D^2\mathbb{P}(\mathrm{SM}_3^m \neq \mathrm{SX}_3^m|\mathrm{SX}_3^m; \theta) = \theta + z$$

In practice, as θ is constrained, we use the logit transformation $\theta = e^\beta/(1 + e^\beta)$ and we can write

$$D^2\mathbb{P}(\mathrm{SM}_3^p = \mathrm{SX}_3^p|\mathrm{SX}_3^p; \beta) = \frac{1}{1 + e^\beta} - \frac{e^\beta}{(1 + e^\beta)^2}z - \frac{e^\beta(1 - e^\beta)}{(1 + e^\beta)^3}z^2$$

and

$$D^2\mathbb{P}(\mathrm{SM}_3^p \neq \mathrm{SX}_3^p|\mathrm{SX}_3^p; \beta) = \frac{e^\beta}{1 + e^\beta} + \frac{e^\beta}{(1 + e^\beta)^2}z + \frac{e^\beta\left(1 - e^\beta\right)}{(1 + e^\beta)^3}z^2.$$

Potentials of the form $D^2\mathbb{P}(\mathrm{SM}_3^m|\mathrm{SX}_3^m; \beta)$ associated with the maternal selector SM_3^m are similarly expressed.

Results. Let us stress the fact that computational shortcuts and in particular *forcing* (see Sections 1.6 and 3.2.3) are crucial in genetic linkage in order to remove unnecessary links. The resulting complexity reduction is more or less spectacular according to family structure and observations as shown for instance in Table 5.3 over four different families proposed in the Mendel package.

Genetic linkage analysis is usually based on a hypothesis testing where a null hypothesis $\theta = 0.5$ is tested against $\theta < 0.5$ using a function of a log-likelihood ratio named the LOD score which is defined as $\mathrm{LOD}(\theta) = \log_{10}\left(L(\theta)/L(0.5)\right)$. $\mathrm{LOD}(\theta)$ is evaluated at $\theta = \hat{\theta}$, an estimate of θ using maximum likelihood estimation (see Section 3.1.3.2).

Defining $Z(\beta) = \log_{10}\left(L\left(e^\beta/(1 + e^\beta)\right)/L(0.5)\right)$, we computed $\mathrm{Z}(\beta)$ for various β and obtained the same values as those computed with the Mendel package for the corresponding $\mathrm{LOD}(\theta)$. The computed derivatives of $\tilde{L}(\beta) = L\left(e^\beta/(1 + e^\beta)\right)$ allow to calculate confidence intervals for θ and to perform likelihood ratio test, Wald test and score test whose results are compared in Table 5.4 for both the KUS family and the whole set of families in `control2a`. As expected, confidence intervals shrink with an increasing number N of individuals. Furthermore the three tests are not equivalent though all p-values are significant. The likelihood ratio is the one

commonly applied in genetic linkage through the LOD score. A further extension of this work could be a comparison of the power of these tests in genetic linkage in different pedigrees.

5.4 Discussion

We introduced in this chapter a novel approach for computing the exact derivatives of the likelihood up to a chosen order d in a parametric BN with parameter $\theta \in \mathbb{R}^p$. Our method is based on the replacement respectively of the conventional product by the Leibniz product and of local probability distributions by derivative generating functions. Its algorithmic complexity is the order of $\mathcal{O}(c \times d^2)$ in time for $p = 1$ and $\mathcal{O}(c \times p^{2d})$ for $p > 1$ where c is the complexity to compute the likelihood itself. Our method can be viewed as an extension of the work of Cowell (1992) and of Nilsson (2001) from moment to derivative generating functions.

Based on Louis identities, Cappé and Moulines (2005) propose an alternative approach using smoothing recursions to compute the last line of Equation (5.4) in the framework of HMMs and $d = 2$. The tree-structure of a junction-tree may allow for a natural extension of their work from HMMs to BNs. On another hand, *formulae* developed by Oakes (1999) and expressed in a BN in Equations (5.2) and (5.3) lead to the computationally demanding quantity $\left[\partial \mathbb{P}(X_{\{\mathrm{pa}(u),u\}}|\mathrm{ev};\theta)/\partial\theta\right]$ and application examples selected by Oakes are restricted to those for which we get a closed *formulae*.

In this context, our main short-term perspectives are the combination and the comparison of different methods. We firstly would like to compare the performances of smoothing recursions proposed by Cappé and Moulines (2005) and the computation of moment generating functions proposed by Cowell (1992) and Nilsson (2001) for obtaining Louis' identities.

Secondly we postulate that, setting order $d = 1$, one can apply our method with an additional outward pass of the sum-product algorithm in order to compute $\left[\partial \mathbb{P}(X_{\{\mathrm{pa}(u),u\}}|\mathrm{ev};\theta)/\partial\theta\right]$ for all u, in linear time in c, offering an extension of Oakes' method to an arbitrary BN when exact computation is feasible. Note additionally that, in the particular choice of maximal order $d = 1$, the Leibniz' product and the conventional product are equivalent, hence one can avoid the modification of the multiplicative law. Combining Oakes' formulae with our method seems to be promising as it could offer a computational complexity reduction over aforementioned methods as well as an easy implementation.

Thirdly the comparison of these methods as well as sensitivity analysis and automatic differentiation (Baydin et al., 2018) in terms of performances still needs to be done and constitute another main perspective of this work.

Family	Bod.+Sto. ($n = 12, \#\text{NA} = 0 + 0$)				Kus. ($n = 22, \#\text{NA} = 0 + 0$)			
	var	edges	τ	complexity	var	edges	τ	complexity
Naive	100	226	10	107,760	200	506	10	518,576
Reducing	100	226	10	3,431	200	506	10	6,410
Forcing	100	226	10	1,946	200	506	10	1,008
New reduction	85	112	5	336	156	159	5	476

Family	Kra. ($n = 27, \#\text{NA} = 5 + 2$)				Neu. ($n = 32, \#\text{NA} = 2 + 2$)			
	var	edges	τ	complexity	var	edges	τ	complexity
Naive	242	602	10	537,176	288	720	10	739,936
Reducing	235	585	10	79,095	284	712	10	12,580
Forcing	235	585	10	75,239	284	712	10	1,698
New reduction	205	360	10	74,292	225	251	5	818

Table 5.3: Time complexity reduction for computing the derivatives of the likelihood up to order $d = 2$ obtained after computational shortcuts detailed in Sections 1.6 and 3.2.3. The reduction is given in terms of number of variables in the selector DAG (var), number of edges in the associated factor graph (edges), treewidth (τ) of an associated junction-tree obtained using the min-fill heuristic and resulting time complexity (complexity). Results are computed for four different families where n denotes the number of family members and # NA denotes the number of unobserved values for the marker + unobserved phenotypes. Naive stands for no other shortcut than pruning.

	$\hat{\boldsymbol{\theta}}$ IC 95%	**LR (p-value)**	**W (p-value)**	**S (p-value)**
KUS(N=22)	$0.059\,[0.008, 0.320]$	$14.574\,(1.3 \times 10^{-4})$	$7.264\,(7.0 \times 10^{-3})$	$32.010\,(1.5 \times 10^{-8})$
ALL (N=93)	$0.193\,[0.106, 0.326]$	$17.012\,(3.7 \times 10^{-5})$	$15.821\,(7 \times 10^{-5})$	$12.900\,(3.3 \times 10^{-4})$

Table 5.4: Confidence intervals for θ and statistics of the likelihood ratio test (LR), Wald test (W) and Score test (S) along with p-values.

Chapter 6

Constrained hidden Markov model for sequence segmentation

Sommaire

6.1 Introduction

6.1.1 Context

The accumulation of sequence data in the last decades, in particular in the field of molecular biology, has raised the need of mathematical and computational tools to extract information out of such large datasets. Biological sequence analyses gather a vast variety of statistical domains. We have previously mentioned word (or pattern) counting in Section 4.1 as an introduction for Chapter 4 for it to be actively using generating functions. The present chapter falls into another field of interest called sequence segmentation which aims at determining segments of homogeneous composition in a sequence, with no focus on particular patterns of interest. Segments of atypical composition may have a biological significance and highlighting them is of interest in a wide range of domains, for instance for detecting transmembrane domains, copy number variation, CpG islands, etc.

There exists a tremendous literature in sequence segmentation with three main types of models (see Auger and Lawrence, 1989; Braun and Muller, 1998, for a review). The first class of models is based on Hidden Markov Models (HMMs) composed of a set of observed variables $X = (X_1, \dots, X_n)$ generated by an underlying hidden state sequence $S = (S_1, \dots, S_n) \in \{1, \dots, L\}^n$ assumed to be a Markov chain. For all $i \in \{1, \dots, n\}$, the emission probability $\mathbb{P}(X_i|S_i)$ is parametrized by $\theta_j \in \{\theta_1, \dots, \theta_L\}$ according to the value taken by S_i. Rabiner (1989) offers a didactic tutorial and Durbin et al. (1998) a detailed development of their applications and many extensions for biological sequences. HMMs are ubiquitous in sequence analysis (Guéguen, 2005; Guédon, 2007) in various fields. Their applications to biological sequence segmentation include in particular the detection of transmembrane domains in proteins (Von Heijne, 1992; Casadio et al., 1996; Rost et al., 1996; Sonnhammer et al., 1998; Krogh et al., 2001), DNA binding sites (Qin et al., 2010; Zaman et al., 2017), copy number variation (Fridlyand et al., 2004; Marioni et al., 2006), CpG islands (Guéguen, 2005) or perform gene prediction (Lukashin and Borodovsky, 1998; Munch and Krogh, 2006).

Note that several variants within the field of biological sequences have been developed for answering related but different questions. For instance pair HMMs, multiple sequence alignment or profile HMM are related to sequence comparison. Such HMM variants fall outside the scope of this thesis (see Durbin et al., 1998; Yoon, 2009, for more details).

Another class of models called change point detection aims at determining abrupt changes in the probability distribution of a stochastic process or a time series. They are classified into *online* and *offline* algorithms. The former process the data as they become available whereas the latter considers an entire dataset. Change point detection is a global optimization problem using a cost function. A quantitative criterion, function of a parametric signal $X = \{X_t\}_{t=1,\dots,n}$ and a set of unknown indexes for break points in the parameter is minimized. The number of break points may also be unknown and to be estimated. Change point detection has been applied to vast variety of domains (quality control, linguistic, music, meteorological data, financial time series, etc.) and in particular in biological sequence analysis, for

instance in the field of copy number variation (Olshen et al., 2004; Picard et al., 2005; Vert and Bleakley, 2010; Niu and Zhang, 2012). See also Hocking et al. (2018) for an R package dedicated to segmentation in genomic sequences. A detailed survey of methods is offered by Aminikhanghahi and Cook (2017) and Truong et al. (2020).

Luong et al. (2013) propose an alternative approach to change point detection methods via a constrained HMM named segment-based HMM by the authors. In contrast, classical HMMs are named level-based HMMs by the authors. The main difference between both types of HMMs layes within the state space of the hidden states such that, in segment-based methods, $S = (S_1, \ldots, S_n) \in \mathcal{S}_n^K$ where $\mathcal{S}_n^K$ is the set of segmentations of K segments in a sequence of length n. Each variable S_i takes value k if the i-th sequence index falls into the k-th segment. The constraint is applied in the choice of the segmentation space, in particular the number of segments K (see also Titsias et al., 2016, for details on an inference in a constrained HMM). This approach allows for taking advantage of the flexibility of the HMM framework with various possible additional evidences of interest to be entered. Moreover, comparing a level-based and a segment-based approach, the former assumes a geometric prior distribution of segment lengths whereas the latter allows for a wider type of priors.

Finally a third class of models including sliding windows and the maximal score are score-based. The maximal score, first defined by Karlin and Altschul (1990) is also commonly called the local score. In brief we consider a sequence of observed iid variables $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ and a scoring function (SF) which assigns a real value called a score to each letter in the alphabet $\mathcal{X}$. The scoring function must satisfy certain properties that will be developed in Section 6.1.3. Both methods are based on the study of the cumulative score of a given segment of a fixed length (sliding windows) or arbitrary length (maximal score). For their easy implementation and interpretation they are actively used in biological sequence analysis (see, for instance Bonhomme et al., 2019, for an example in the detection of quantitative trait loci). The choice of a SF is currently made according to the context and the feature of interest. A vast variety of SFs has been developed from empirical knowledge and/or supervised learning based on various properties of interest, for instance amino acid charges, hydrophobicity, size, weight, etc. (see Karlin and Altschul, 1990; Karlin, 2005, for a selection of examples). Among several other examples, the Expasy ProtScale website (`https://web.expasy.org/protscale/`) provides 57 amino-acid SFs for the study of various features in proteins. One can for instance download the scale based on hydrophobic/hydrophilic properties of amino acids developed by Kyte and Doolittle (1982) and later refined by Zhao and London (2006) with statistical analyses of databases of known soluble and transmembrane proteins. In the field of evolution and in particular for detecting candidate loci for positive selection, Grossman et al. (2013) used functions of appropriate multiple signal tests, Enard et al. (2014) used the correlation between neutral diversity and recombination rate measured in 500-kb windows and Fariello et al. (2017) developed a function of $-\log_{10}(p_i)$ where p_i is the p-value of a statistical test aiming at rejecting the null hypothesis of neutral evolution at position i on the genome.

Score based methods like sliding windows or the maximal score share similarities with scan statistics in the sense that, in scan statistics, events, usually modeled by

a Bernoulli or a Poisson process, are assigned a score. Its main focus is related to the distribution of the number of events in windows of fixed size. Scan statistics have been introduced by Naus (1963) and actively developed since then (Glaz et al., 2009; Chen and Glaz, 2016; Zhao and Glaz, 2016, 2017). They have been applied to a variety of fields and in particular to DNA sequences (Takai and Jones, 2002; Zhang et al., 2016; Pellin and Di Serio, 2016; He et al., 2019; Guo et al., 2021). Scan statistics fall outside the scope of this thesis.

In this context, the purpose of the work developed in the present chapter is double. 1) To the best of our knowledge, all existing SFs are learnt empirically and/or with supervised methods. Under certain conditions, Mercier and Nuel (2021) proved that there exists a dual relation between score-based methods and a constrained HMM. Our first goal is to take advantage of this dual relation in order to develop an unsupervised method for the statistical learning of a SF for the maximal score with confidence intervals. The underlying model is a constrained segment-based HMM for it to be able to constraint the number of atypical segments in a sequence as well as constraint two non-adjacent segments to share common emission probabilities. 2) Secondly, we will extend the maximal score to the search of more than one atypical segment and/or segments of multiple types with an arbitrary prior distribution for the number of segments. The part related to the equivalence between score-based methods and a constraint HMM is done in collaboration with Sabine Mercier (University of Toulouse) and the part concerning the extension to an arbitrary prior for the number of segments is realized in collaboration with Vittorio Perduca (University Paris Descartes).

This chapter is divided into several sections and subsections and starts with two introductions (Sections 6.1.2 and 6.1.3) respectively to level-based HMMs for biological sequence segmentation and to the maximal score.

Section 6.2 is dedicated to our first purpose, the unsupervised SF learning, and starts in Section 6.2.1 with the presentation of main results in (Mercier and Nuel, 2021) and of the constrained segment-based HMM later used. In Section 6.2.2, we develop the method and the expression of quantities needed for parameter estimation along with confidence intervals. In Section 6.2.3 we propose a selection of other inferences of interest related to the existence and the localization of an atypical segment. Section 6.2.4 is an applied section with simulated datasets and real datasets of transmembrane proteins downloaded from the UniProt database[1] in order to illustrate the interest of the method as well as its weaknesses.

Section 6.3 is dedicated to our second goal, the extension of the maximal score to multiple segments of same type and multiple segment types. We start in Section 6.3.1 by detailing the extension of the segment-based HMM using polynomial potentials in order to allow for an arbitrary prior distribution of the number of segments. We focus in particular on two quantities of interest: the computation of the posterior distribution of the number of segments and the marginal posterior probabilities of individual states in sequences containing multiple segments. In Section 6.3.2, we will see a limited selection of application examples over simulated datasets.

Finally in Section 6.4 we recall main advantages and disadvantages of the method

[1] `http://www.uniprot.org/uniprot/`

and present future perspectives.

6.1.2 Hidden Markov models in biological sequence analysis

In this section inspired by Rabiner (1989); Durbin et al. (1998); Yoon (2009) we recall principles of (level-based) HMMs applied to the detection of segments of atypical composition in a sequence using notation introduced in Chapter 1 and in particular in the Section dedicated to the particular case of HMMs (Section 1.7.3). We consider a set of observed variables $X = \{X_1, \dots, X_n\} \in \mathcal{X}^n$ generated by an underlying state sequence $S = (S_1, \dots, S_n)$ where each S_i takes its values in $\mathcal{S} = \{1, \dots, L\}$ and denotes the underlying state of the i^{th} observation X_i. The space $\mathcal{X}$ is usually called the alphabet and can, for instance, be the set of nucleic acids $\{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}$ for a DNA sequence or the set of the 20 amino acids for a protein to study. We denote by x_i the observed value taken by X_i. The initial state S_1 takes value $s \in \mathcal{S}$ with *initial state probability* $\mu(s)$. We assume that S is a first order Markov chain, i.e. for all $i \in \{2, \dots, n\}$, $\mathbb{P}(S_i|S_1, \dots, S_{i-1}) = \mathbb{P}(S_i|S_{i-1})$ and, for all $r, s \in \mathcal{S}$, $\mathbb{P}(S_i = s|S_{i-1} = r) = \pi_i(r, s)$ where π_i is called the *transition probability* at position i. If $\pi_i = \pi_j$ for all $i, j \in \{2, \dots, n\}$, the model is said to be homogeneous. The probability of observing $X_i = x_i$ depends only on S_i such that $\mathbb{P}(X_i|S) = \mathbb{P}(X_i|S_i)$ and $\mathbb{P}(X_i = x_i|S_i = s) = \eta(s, x_i)$ where η is called the *emission probability* of x_i at state s. The parameter θ is the set of values of probability measures such that $\theta = \{\mu(s), \{\pi_i(r, s)\}_{i=2,\dots,n}, \{\eta(s, x_i)\}_{i=1,\dots,n}\}$ for all $r, s \in \mathcal{S}$. In order to simplify the notation, let us assume that, for all $s \in \mathcal{S}$, $\mu(s) = \pi_1(1, s)$. Therefore, the joint probability $\mathbb{P}(X, S|\theta)$ is given by

$$\mathbb{P}(X, S) = \prod_{i=1}^{n} \pi_i(S_{i-1}, S_i)\eta(S_i, X_i)$$

where $S_0 = 1$ by convention.

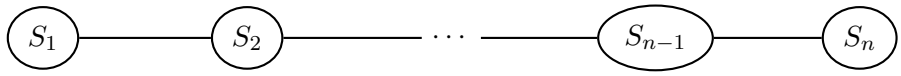
In this section, we detail the sum-product algorithm, also called Forward-Backward algorithm (FB algorithm) in HMMs seen in Chapter 1, Section 1.7.3, over the HMM introduced above. Despite some redundancies, this choice is motivated by the desire to render the present chapter independent for future publication and by the sake of clarity as HMMs do not require the entire set of tools seen in Chapter 1. In particular, factor graphs are trees, hence the choice of an elimination ordering is trivial to obtain junction-trees which are sequences.

Let $\text{ev} = \{X_i \in \mathcal{X}_i^* \subseteq \mathcal{X}\}$, we introduce, for all $i \in \{1, \dots, n\}$, the potential

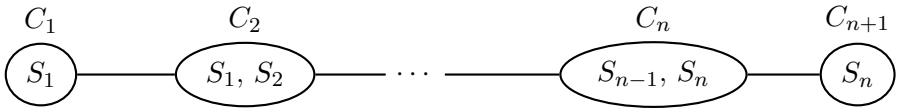
$$\phi_i(S_{i-1}, S_i) = \pi_i(S_{i-1}, S_i) \sum_{x_i \in \mathcal{X}_i^*} \eta(S_i, x_i) \tag{6.1}$$

and consequently we have

$$\mathbb{P}(\text{ev}, S|\theta) = \prod_{i=1}^{n} \phi_i(S_{i-1}, S_i). \tag{6.2}$$



(a) Factor graph H_ϕ



(b) Junction-tree J defined by $\mathrm{VE}(\phi, S, \sigma)$

Figure 6.1: Factor graph H_ϕ (top) where ϕ is the set of potentials defined in Equation (6.1) and junction-tree J (bottom) defined by $\mathrm{VE}(\phi, S, \sigma)$ where $\sigma = (S_1, \ldots, S_n)$. For all $i \in \{1, \ldots, n\}$, $C_i^* = \phi_i$ is the potential injected in C_i.

Let $\phi = \{\phi_i\}_{i=1,\ldots,n}$ be the set of potentials and $\sigma = (S_1, \ldots, S_n)$ be a perfect elimination ordering over the factor graph H_ϕ, Figure 6.1 is a graphical representation of H_ϕ and the junction-tree J defined by $\mathrm{VE}(\phi, S, \sigma)$.

Three main quantities are computed in practice: the likelihood of the model, marginal individual posterior state probabilities and the most probable state sequence. For the sake of simplicity, we restrict this chapter to the evidence $\mathrm{ev} = \{X = x\}$ where $x = (x_1, \ldots, x_n)$ is the vector of observed values taken by X. However, more complexe evidences can be similarly considered.

Likelihood. The likelihood of the model is given by

$$\mathcal{L}(\theta) \stackrel{\text{def}}{=} \mathbb{P}(X = x|\theta) = \sum_{S \in \mathcal{S}^n} \mathbb{P}(X = x, S|\theta)$$

for which a brute force computation is of the order of $\mathcal{O}(L^n)$ in time. As seen in Section 1.7.3, the FB algorithm allows for a complexity reduction. For the likelihood of the model, the forward pass is sufficient. We define *forward* messages to be, for all $i \in \{1, \ldots, n\}$,

$$F_i(S_i) \stackrel{\text{def}}{=} \mathbb{P}(X_1 = x_1, \ldots, X_i = x_i, S_i|\theta). \tag{6.3}$$

By definition we have $F_1(S_1) = \phi_1(1, S_1)$ and for all $i \in \{2, \ldots, n\}$,

$$F_i(S_i) = \sum_{S_1, \ldots, S_{i-1}} \prod_{j=1}^{i} \phi_j(S_{j-1}, S_j)$$

which can be recursively computed with initialization, for all $s \in \mathcal{S}$, $F_1(s) = \phi_1(1, s)$ and for $i = 2, \ldots, n$,

$$F_i(s) = \sum_{r \in \mathcal{S}} F_{i-1}(r)\phi_i(r, s). \tag{6.4}$$

Applying Theorem 2 over the last separator and summing S_n out of the resulting potential, one can compute the likelihood $\mathbb{P}(X = x|\theta) = \sum_{s \in \mathcal{S}} F_n(s)$ with a total time complexity of the order of $\mathcal{O}(n \times L^2)$.

Individual posterior state. Performing an additional backward pass of same order time complexity, one can compute the marginal posterior state probability of any hidden state S_i. We define the backward messages to be, for all $i \in \{1, \dots, n\}$,

$$B_i(S_i) \overset{\text{def}}{=} \mathbb{P}(X_{i+1} = x_{i+1}, \dots, X_n = x_n | S_i; \theta). \tag{6.5}$$

Note that $B_n(S_n) = 1$ and for $i \in \{1, \dots, n-1\}$,

$$B_i(S_i) = \sum_{S_{i+1}, \dots, S_n} \prod_{j=i+1}^{n} \phi_j(S_{j-1}, S_j).$$

Therefore, backward messages can be recursively computed with initialization, for all $s \in \mathcal{S}$, $B_n(s) = 1$ and for $i = n, \dots, 2$, for all $r \in \mathcal{S}$,

$$B_{i-1}(r) = \sum_{s \in \mathcal{S}} \phi_i(r, s) B_i(s).$$

Finally, applying Theorem 2 over a chosen separator S_i, we have $\mathbb{P}(S_i = s, X = x|\theta) = F_i(s)B_i(s)$. Therefore, the marginal posterior probability of S_i is given by

$$\mathbb{P}(S_i = s | X = x; \theta) = \frac{1}{\mathcal{Z}} F_i(s) B_i(s) \tag{6.6}$$

where the partition function $\mathcal{Z} = \mathbb{P}(X = x|\theta) = \sum_{r \in \mathcal{S}} F_i(r)B_i(r)$. Note that $\mathcal{Z}$ is a constant in i and can be computed with a choice of any index $i \in \{1, \dots, n\}$.

Underflow and log computations. Let us stress the fact that, in practice, computational tricks for avoiding underflow issues may be inevitable for long sequences. One may apply, for instance, a recursive rescale of forward and backward messages detailed in Section 1.5. We briefly adapt the method proposed in Section 1.5 for keeping quantities recursively computed in a tractable range. Let us define rescaled forward messages and forward logarithmic factors to be respectively, for all $i \in \{1, \dots, n\}$ and $s \in \mathcal{S}$, $\tilde{F}_i(s) = F_i(s) / \sum_{r \in \mathcal{S}} F_i(r)$ and $L_i = \log \sum_{r \in \mathcal{S}} F_i(r)$. These quantities can be computed in linear in n time complexity with the FB algorithm with an additional rescaling at each step. The process is detailed thereafter where the quantity aux is updated at each step. The algorithm is initialized with $\text{aux} = \sum_{s \in \mathcal{S}} F_1(s) = \sum_{s \in \mathcal{S}} \phi_1(1, s)$, $L_1 = \log(\text{aux})$ and for $s \in \mathcal{S}$, $\tilde{F}_1(s) = \phi_1(1, s)/\text{aux}$. Then, for $i = 2, \dots, n$, $\text{aux} = \sum_{r,s \in \mathcal{S}} \tilde{F}_{i-1}(r)\phi_i(r, s)$, for $s \in \mathcal{S}$

$$\tilde{F}_i(s) = \frac{1}{\text{aux}} \sum_{r \in \mathcal{S}} \tilde{F}_{i-1}(r)\phi_i(r, s) \qquad \text{and} \qquad L_i = L_{i-1} + \log(\text{aux}).$$

Multiplying indeed both numerator and denominator by $\exp(L_{i-1})$ in the first equation proves it. Furthermore, we have $L_i = \log\left(\sum_{r,s \in \mathcal{S}} \left[\tilde{F}_{i-1}(r) \exp(L_{i-1})\right] \phi_i(r, s)\right)$ which proves the second equation.

Similarly backward rescaled messages and logarithmic factors are defined, for all $i \in \{1, \dots, n\}$, for $s \in \mathcal{S}$, by $\tilde{B}_i(s) = B_i(s) / \sum_{r \in \mathcal{S}} B_i(r)$ and $M_i = \log \sum_{s \in \mathcal{S}} B_i(s)$

and recursively computed with initialization, for $s \in \mathcal{S}$, $\tilde{B}_n(s) = 1/n$ and $M_n = \log(n)$. Then for $i = n, \ldots, 2$, aux $= \sum_{r,s\in\mathcal{S}} \phi_i(r,s)\tilde{B}_i(s)$ and for all $r \in \mathcal{S}$,

$$\tilde{B}_{i-1}(r) = \frac{1}{\text{aux}} \sum_{s\in\mathcal{S}} \phi_i(r,s)\tilde{B}_i(s) \qquad \text{and} \qquad M_{i-1} = M_i + \log(\text{aux}).$$

Finally we have

$$\log(\mathcal{Z}) = \log \mathbb{P}(X = x|\theta) = \log\left(\sum_{s\in\mathcal{S}} \tilde{F}_i(s)\tilde{B}_i(s)\right) + L_i + M_i$$

which is a constant in index i. If the log-likelihood is solely needed, one can avoid the backward pass and compute $\log \mathbb{P}(X = x|\theta)$ choosing $i = n$ in the above equation. Moreover posterior state probabilities are given by

$$\mathbb{P}(S_i = s|X = x;\theta) = \frac{\tilde{F}_i(s)\tilde{B}_i(s)}{\sum_{r\in\mathcal{S}} \tilde{F}_i(r)\tilde{B}_i(r)}.$$

Multiplying indeed both numerator and denominator by $\exp(L_i + M_i)$ we obtain Equation (6.6).

Most probable state sequence. The most probable state sequence is defined as

$$S^* \stackrel{\text{def}}{=} \text{MAP}(S|X = x;\theta) \stackrel{\text{def}}{=} \arg\max_{S\in\mathcal{S}^n} \mathbb{P}(S|X = x;\theta) = \arg\max_{S\in\mathcal{S}^n} \mathbb{P}(X = x, S|\theta) \quad (6.7)$$

for which a brute force computation is also exponential in n but one can drop it to linear in n with the max-product (or max-sum) algorithm as mentioned in Section 1.8. The max-product algorithm is called the Viterbi algorithm (Viterbi, 1967) in the framework of HMMs (see Forney, 1973, for a comprehensive tutorial). A small adaptation of the Viterbi algorithm is proposed in (Schwartz and Chow, 1990) and (Nilsson and Goldberger, 2001) for inferring the M most probable state sequences. We detail thereafter the Viterbi algorithm in its max-product version for it to be the mirror version of the sum-product algorithm but it can be advantageously replaced by the max-sum algorithm over log-potentials in order to avoid a computational underflow.

We define the max-forward messages to be, for all $i \in \{1, \ldots, n\}$,

$$F_i^{\max}(S_i) \stackrel{\text{def}}{=} \max_{S_1,\ldots,S_{i-1}} \mathbb{P}(X_1 = x_1, \ldots, X_i = x_i, S_1, \ldots, S_i|\theta).$$

Note that for $i \in \{2, \ldots, n\}$, $F_i^{\max}(S_i) = \max_{S_1,\ldots,S_{i-1}} \prod_{j=1}^{i} \phi_j(S_{j-1}, S_j)$ with $F_1^{\max}(S_1) = \phi_1(1, S_1)$. Therefore, one can recursively compute max-forward messages over the JT J represented in Figure 6.1b using the same recursion as the one developed for the likelihood and replacing the $\sum$ by the max operator. At the end of the recursion, as $\max_{s\in\mathcal{S}} F_n^{\max}(s) = \max_{S\in\mathcal{S}^n} \mathbb{P}(X = x, S|\theta)$, the most probable state sequence S^* is traced back from $\max_{s\in\mathcal{S}} F_n^{\max}(s)$ for a total time complexity of the order of

$\mathcal{O}(n \times |\mathcal{S}|^2)$. Alternatively, one can define max-backward messages to be, for all $i \in \{1, \dots, n\}$,

$$B_i^{\max}(S_i) \stackrel{\text{def}}{=} \max_{S_{i+1},\dots,S_n} \mathbb{P}(X_{i+1} = x_{i+1}, \dots, X_n = x_n, S_{i+1}, \dots, S_n | S_i; \theta), \tag{6.8}$$

which can be recursively computed with a backward recursion over J with the algorithm developed for marginal individual posterior state probabilities where the $\sum$ is replaced by the max operator. At the end of the recursion, the most probable state sequence is given at any position $i \in \{1, \dots, n\}$ by $\arg\max_{s \in \mathcal{S}} F_i^{\max}(s) B_i^{\max}(s)$.

Parameter estimation. Main methods used for optimizing the likelihood of such incomplete models include the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) previously developed in the particular case of HMMs in Baum et al. (1970) or gradient-based methods which again take advantage of the FB algorithm for a complexity reduction. Detailed explanations are provided in Rabiner (1989) or Cappé et al. (2005). When an exact computation of the likelihood is intractable, approximate method may be required such as approximate Bayesian computations (Marin et al., 2012; Dean et al., 2014).

6.1.3 Maximal score and maximal scoring segment

The maximal score was introduced by Karlin and Altschul (1990) and is appreciated for biological sequence analysis for it to be easy to implement by dynamic programming and to interpret. Let $X = (X_1, \dots, X_n) \in \mathcal{X}^n$ be a sequence of n observed letters, a scoring function (SF) $f : \mathcal{X} \to \mathbb{R}$ is a real-valued function which assigns a real number called score to each letter in the alphabet according to the feature of interest (for instance, $f(\texttt{A}) = f(\texttt{T}) = -2$ and $f(\texttt{C}) = f(\texttt{G}) = 1$ for the search of a GC-rich region in a DNA sequence). The *maximal score* is defined as

$$H_f(X) \stackrel{\text{def}}{=} \max_{I \in \mathcal{I}} \sum_{k \in I} f(X_k) \tag{6.9}$$

where $\mathcal{I}$ is the set of segments (including the empty one) in sequences of length n and the sum over the empty set is null by convention. Note that consequently, for any sequence X, we have $H_f(X) \geq 0$.

The *maximal scoring segment* is defined as one segment (possibly the empty one) that realizes the maximal score:

$$I_f(X) \stackrel{\text{def}}{=} \arg\max_{I \in \mathcal{I}} \sum_{k \in I} f(X_k). \tag{6.10}$$

The *Lindley process* (Lindley, 1952) allows for a dynamic programming of the maximal scoring segment with initialization $H_0 = 0$, then for $i = 1, \dots, n$, $H_i = \max(0, H_{i-1} + f(X_i))$, we get $H_f(X) = \max_i H_i$. The result $(H_1, \dots, H_n)$ being several excursions above zero, a segment $I_f(X) = [i_{\text{start}}, i_{\text{stop}}]$ is defined as $i_{\text{stop}} = \arg\max_{i \in \{1,\dots,n\}} H_i$ and $i_{\text{start}} = 1 + \max_{i < i_{\text{stop}}; H_i = 0}$.

6.2 Unsupervised learning of a scoring function for the maximal score with a constrained hidden Markov model

6.2.1 Relation between score-based methods and hidden Markov models

The local score aims at localizing the (or one) maximal scoring segment as defined in Equation (6.10) with a chosen SF f. However, given a SF, the maximal score does not allow for capturing suboptimal segments of interest. Let us illustrate this point with a graphical representation of a simple example in Figure 6.2. The maximal score is computed for two simulated sequences of length 1200 letters each, over the alphabet of nucleic acids $\mathcal{X} = \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$. Both sequences contain a high GC-content region delimited by the greyed region. In both simulation schemes, letters are assumed to be independent and identically distributed (iid) and generated with a uniform distribution of the four nucleic acids outside the greyed region. Inside the greyed region, letters are iid and we used a GC-ratio of 7/3 for the first sequence (Figure 6.2a) and 4.5/3 for the second sequence (Figure 6.2b). As explained in Karlin and Altschul (1990), a chosen SF for applying the maximal score must satisfy the following two conditions:

$$\mathbb{E}_{H_0}\big[f(X_i)\big] < 0 \tag{6.11}$$

where $\mathbb{E}_{H_0}$ is the expectation with the background (non-atypical) model and

$$\exists i \in \{1, \dots n\} \text{ such that } f(x_i) > 0. \tag{6.12}$$

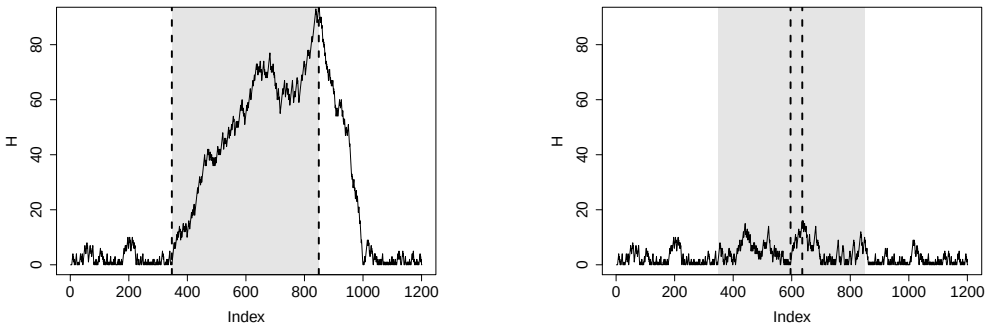
We choose the SF $f : \mathcal{X} \mapsto \mathbb{R}$, $f(\mathtt{A}) = f(\mathtt{T}) = -2$ and $f(\mathtt{C}) = f(\mathtt{G}) = +1$ that verifies both Assumptions 6.11 and 6.12. The maximal scoring segment as defined in Equation (6.10) and computed with the Lindley process is graphically delimited by the two dashed lines. Note that it accurately covers the true high GC-content segment when applied to a sequence of high GC-ratio between atypical and non-atypical segments (Figure 6.2a) but it misses several suboptimal segments in the sequence of a lower ratio (Figure 6.2b).

In order to overcome that limitation, Mercier and Nuel (2021) proposed a novel approach based on the following Gibbs measure:

$$\forall I \in \mathcal{I}, \qquad \mathbb{P}^{f,T}_{\text{Gibbs}}(I|X) \propto \exp\left(\frac{1}{T}\sum_{k\in I} f(X_k)\right) \tag{6.13}$$

where $\mathcal{I}$ is the set of segments (including the empty one) in sequences on length n and $T > 0$, called the temperature of the system, is a contrast parameter. Indeed $\mathbb{P}^{f,T}_{\text{Gibbs}}(I|X)$ tends towards a uniform distribution as $T \to +\infty$ and towards a Dirac distribution in the segment of minimal energy as $T \to 0$. The quantity $\big\{-\sum_{k\in I} f(X_k)\big\}$ is called the energy of the segment $I \in \mathcal{I}$. Hence a maximal scoring segment (defined in Equation 6.10) is a segment of minimal energy.

The authors introduced the (hidden) state sequence $S = (S_1, \dots, S_n) \in \mathcal{S}_n^{\{0,1\}}$ where $\mathcal{S}_n^{\{0,1\}}$ is the set of sequences of length n containing zero or one segment.



(a) GC-ratio of 7/3 versus 1 in atypical versus non-atypical segments.

(b) GC-ratio of 4.5/3 versus 1 in atypical versus non-atypical segments.

Figure 6.2: Graphical representation of the maximal score for two different simulated sequences with one GC-rich segment highlighted by the greyed region. Both datasets are simulated with a GC-ratio of 1 outside the atypical segment and 7/3 (respectively 4.5/3) inside the atypical segment for the sequence on the left (respectively on the right). The maximal scoring segment computed with the Lindley process using the SF $f : \mathcal{X} \mapsto \mathbb{R}$, $f(\mathtt{A}) = f(\mathtt{T}) = -2$ and $f(\mathtt{C}) = f(\mathtt{G}) = +1$ is delimited by the two dashed lines.

Each S_i takes its values in $\{1, 2, 3\}$ and denotes the position of index i regarding the atypical segment such that $S_i = 1$ (respectively $S_i = 2$ and $S_i = 3$) if the i^{th} index is before (respectively inside and after) the atypical segment. Hence we have $\mathcal{S}_n^{\{0,1\}} = \{S = \{1, 2, 3\}^n; S_i - S_{i-1} \in \{0, 1\} \text{ for } i = 1, \dots, n\}$ with $S_0 = 1$ by convention. Note that there exists a bijection between $\mathcal{I}$ and $\mathcal{S}_n^{\{0,1\}}$ such that $\forall I \in \mathcal{I}, \exists! S \in \mathcal{S}_n^{\{0,1\}}$ and $\forall S \in \mathcal{S}_n^{\{0,1\}}, \exists! I \in \mathcal{I}$ such that $I = \{i \in \{1, \dots, n\}; S_i = 2\}$. Let us highlight the particular case of the empty segment wich corresponds to the sequence $(S_i = 1)_{i=1,\dots,n}$. Let $q_0(.)$ and $q_1(.)$ be two multinomial distributions, the authors introduced the following generative model:

$$\mathbb{P}_{\text{GM}}^{q_0,q_1}(X|S) = \prod_{i=1}^{n} q_0(X_i)^{\mathbb{1}_{\{S_i \neq 2\}}} q_1(X_i)^{\mathbb{1}_{\{S_i=2\}}} \tag{6.14}$$

where $\mathbb{1}_{\{.\}}$ is the indicator function and for all $a \in \mathcal{X}$, $q_0(a)$ (respectively $q_1(a)$) is the emission probability of a outside (respectively inside) the atypical segment, i.e., for all $i \in \{1, \dots, n\}$, $q_0(X_i) = \mathbb{P}_{\text{GM}}^{q_0,q_1}(X_i|S_i \neq 2)$ and $q_1(X_i) = \mathbb{P}_{\text{GM}}^{q_0,q_1}(X_i|S_i = 2)$. Let $\mathcal{M}_{\mathcal{X}}$ be the set of multinomial distributions over $\mathcal{X}$, as proved in Mercier and Nuel (2021), we have the following two theorems:

Theorem 5. *$\forall (q_0, q_1) \in \mathcal{M}_{\mathcal{X}}^2$, $\exists! \sigma : \mathcal{X} \to \mathbb{R}$ such that, $\forall S \in \mathcal{S}_n^{\{0,1\}}$, $\mathbb{P}_{\text{Gibbs}}^{\sigma,1}(S|X) = \mathbb{P}_{\text{GM}}^{q_0,q_1}(S|X)$ and $\forall a \in \mathcal{X}$, $\sigma(a) = \log \frac{q_1(a)}{q_0(a)}$.*

Theorem 6. *$\forall q_0 \in \mathcal{M}_{\mathcal{X}}$ and $\forall f : \mathcal{X} \to \mathbb{R}$ verifying Assumptions 6.11 and 6.12, $\exists! T > 0$ and $q_1 \in \mathcal{M}_{\mathcal{X}}$, such that $\forall S \in \mathcal{S}_n^{\{0,1\}}$, $\mathbb{P}_{\text{GM}}^{q_0,q_1}(S|X) = \mathbb{P}_{\text{Gibbs}}^{f,T}(S|X)$ and $\forall a \in \mathcal{X}$, $q_1(a) = q_0(a) \exp \frac{f(a)}{T}$. Note that $\mathbb{P}_{\text{GM}}^{q_0,q_1}(S|X) = \mathbb{P}_{\text{Gibbs}}^{\sigma,1}(S|X)$ with $\sigma = \log \frac{q_1}{q_0} = \frac{f}{T}$.*

Therefore the two models $\mathbb{P}_{\text{GM}}^{q_0,q_1}$ and $\mathbb{P}_{\text{Gibbs}}^{\sigma,1}$ are equivalent with $\sigma = \log(q_1/q_0) = f/T$. Note that Karlin and Altschul (1990) proved that the SF $\log(q_1/q_0)$ always verifies Assumptions 6.11 and 6.12.

In order to make the notation lighter, we will from now on denote the distribution $\mathbb{P}_{\text{GM}}^{q_0,q_1}$ simply by $\mathbb{P}$. Furthermore we consider $\mathcal{X}$ to be $\mathcal{X} = \{1, \dots, d\}$ and we denote by $q = \big((q_0(a))_{a=1,\dots,d}, (q_1(a))_{a=1,\dots,d}\big)$, the vector of letter frequencies in non-atypical and atypical segments.

We now introduce the HMM used for future inferences. For all $s \in \{1,2,3\}$, for all $a \in \mathcal{X}$, we define the emission probabilities to be $\nu(s,a) = q_0(a)^{\mathbb{1}_{s\neq 2}} q_1(a)^{\mathbb{1}_{s=2}}$. Assuming a uniform distribution over $\mathcal{S}_n^{\{0,1\}}$, we define the transition function $\tau : \{1,2,3\}^2 \to \{0,1\}$ such that, for all $r, s \in \{1,2,3\}$, $\tau(r,s) = \mathbb{1}_{\{s-r\in\{0,1\}\}}$ where $\mathbb{1}_{\{.\}}$ is the indicator function. Hence we can write

$$\mathbb{P}(S, X = x|q) \propto \prod_{i=1}^{n} \tau(S_{i-1}, S_i)\eta_i(S_i) \tag{6.15}$$

where $S_0 = 1$ by convention and each η_i is defined over $\{1,2,3\}$ such that, for all $s \in \{1,2,3\}$, $\eta_i(s) = \nu(s, x_i)$. Note that τ is a simple transition function which ensures a uniform distribution over the chosen state sequence space $\mathcal{S}_n^{\{0,1\}}$, not proper transition probabilities, and the quantity on the right is solely proportional to $\mathbb{P}(S, X = x|q)$.

Interestingly, for a fixed parameter q, dividing Equation (6.15) by $\prod_{i=1}^{n} q_0(x_i)$ we can also write

$$\mathbb{P}(S, X = x|q) \propto \prod_{i=1}^{n} \tau(S_{i-1}, S_i) \left(\frac{q_1(x_i)}{q_0(x_i)}\right)^{\mathbb{1}_{\{S_i=2\}}} \tag{6.16}$$

and therefore the maximal scoring segment is equivalently defined by the maximal score using SF $\sigma = \log(q_1/q_0)$ or the HMM defined in Equation (6.16) using the Viterbi or max-sum algorithm (see Equation 6.7) by

$$I_\sigma(X) = \arg\max_{I\in\mathcal{I}} \sum_{k\in I} \log\frac{q_1(x_k)}{q_0(x_k)} = \arg\max_{I\in\mathcal{I}} \sum_{k\in I} \sigma(x_k).$$

6.2.2 Parameter estimation

Let N be the number of atypical segments in a sequence, in this section we detail the expression and the practical computation of quantities needed for the estimation of $\sigma = \log(q_1/q_0)$ conditional on $N = 1$ and for the estimation of the second order derivatives of the log-likelihood as well as our assumptions for obtaining confidence intervals (CIs) for the learnt SF. We applied the EM algorithm for the estimation of σ and the method proposed by Lefebvre and Nuel (2018) and detailed in Chapter 5 for the estimation of second order derivatives of the log-likelihood. For the rest of the chapter, forward and backward messages are functions of q, however, for the sake of simplicity, we will drop q in the notation whenever there is no confusion. In practice, all messages are implemented in their rescaled and logarithmic versions (see Section 6.1.2, paragraph "Underflow and log computations"), however and for the sake of readability, explanations are given with their non-rescaled version.

Likelihood and maximization. With no conditioning on the number of segments, the likelihood of q is given by

$$\mathcal{L}(q) = \mathbb{P}(X = x|q) = \sum_{S \in \mathcal{S}_n^{\{0,1\}}} \mathbb{P}(S, X = x|q) \propto \prod_{i=1}^{n} \tau(S_{i-1}, S_i)\eta_i(S_i)$$

as $\tau \propto \mathbb{P}$. Note that the quantity on the right side of the equation can be recursively computed with a forward pass of the FB algorithm over J (see Figure 6.1b) with the set of potentials $\phi = \{\phi_i\}_{i=1,\dots,n}$ where for all $i \in \{1, \dots, n\}$, $\phi_i = \tau\eta_i$. Let us denote by $F_1, \dots, F_n$, the n forward messages recursively computed during that forward pass.

Noticing that $\{N = 1\} \equiv \{S_n \in \{2, 3\}\}$, the likelihood of the model conditional on $N = 1$ segment is given by

$$\mathcal{L}(q|S_n \in \{2, 3\}) = \frac{\mathbb{P}(X = x, S_n \in \{2, 3\}|q)}{\mathbb{P}(S_n \in \{2, 3\})} \propto \mathbb{P}(X = x, S_n \in \{2, 3\}|q) \tag{6.17}$$

as $\mathbb{P}(S_n \in \{2, 3\})$ is a constant in q. Hence, we get $\mathcal{L}(q|S_n \in \{2, 3\}) \propto \sum_{s \in \{2,3\}} F_n(s)$ in time complexity of the order of $\mathcal{O}(n)$.

Out of classical optimization algorithms applicable for maximizing $\mathcal{L}(q|S_n \in \{2, 3\})$, we used the EM algorithm for which we need to compute, for a given parameter q, $\mathbb{P}(S|X = x, S_n \in \{2, 3\}; q) = \mathbb{P}(S|\text{ev}; q)$ with $\text{ev} = \{X = x, S_n \in \{2, 3\}\}$. Entering ev in the HMM, one can add a backward pass of the FB algorithm over J with potentials $\tilde{\phi} = \{\tilde{\phi}_i\}_{i=1,\dots,n}$ where, for all $i \in \{1, \dots, n-1\}$, $\tilde{\phi}_i = \phi_i$ and $\tilde{\phi}_n$ is defined on $\{1, 2, 3\}$ such that $\tilde{\phi}_n(1) = 0$ and, for $s \in \{2, 3\}$, $\tilde{\phi}_n(s) = \phi_n(s)$. We denote by $B_n, \dots, B_1$ the set of backward messages recursively computed during that pass in linear in n time complexity.

Remark 1: In practice and equivalently, the set of potentials ϕ remains unchanged and the backward pass is initialized with $B_n(1) = 0$ and for $s \in \{2, 3\}$, $B_n(s) = 1$. Thus τ ensures a uniform distribution over the chosen state space $\mathcal{S}_n^{\{1\}}$, the set of sequences containing exactly one segment.

Remark 2: The HMM is said to be constrained as we impose $N = 1$ segment and we constraint the transition matrix.

Finally quantities needed are given by

$$\mathbb{P}(S_i = s|X = x, S_n \in \{2, 3\}; q) = \frac{1}{\mathcal{Z}} F_i(s)B_i(s) \tag{6.18}$$

where

$$\mathcal{Z} = \sum_{s \in \{1,2,3\}} F_j(s)B_j(s) \propto \mathcal{L}(q|S_n \in \{2, 3\}) \tag{6.19}$$

is a constant in j, hence it can be computed choosing any position $j \in \{1, \dots, n\}$.

Derivatives. As q is constrained, we introduce an unconstrained parameter $\theta = (\theta_1^0, \dots, \theta_{d-1}^0, \theta_1^1, \dots, \theta_{d-1}^1) \in \mathbb{R}^{2(d-1)}$ such that, for $k \in \{0, 1\}$, for $a \in \mathcal{X} = \{1, \dots, d\}$, $q_k(a) = q_k^\theta(a) \propto \exp(\theta_a^k)$ with $\theta_d^0 = \theta_d^1 = 0$ by convention. We denote by q^θ, the vector

$((q_0^\theta(a))_{a\in\mathcal{X}}, (q_1^\theta(a))_{a\in\mathcal{X}})$. Let $\hat{\theta} = (\hat{\theta}_1^0, \ldots, \hat{\theta}_{d-1}^0, \hat{\theta}_1^1, \ldots, \hat{\theta}_{d-1}^1) = \arg\max_\theta \mathcal{L}(\widehat{q^\theta}|S_n \in \{2,3\})$, we assume that $\theta \sim \mathcal{N}(\hat{\theta}, \hat{\Sigma})$. The matrix $\hat{\Sigma}$ is estimated with the observed Fisher information matrix at $\hat{\theta}$, $I_F(\hat{\theta}) = -\nabla^2 \log \mathcal{L}(\widehat{q^\theta}|S_n \in \{2,3\})$ such that we assume that $\hat{\Sigma} \approx I_F(\hat{\theta})^{-1}$. We used the method proposed by Lefebvre and Nuel (2018) for the estimation of $I_F(\hat{\theta})$. Note that one could chose another alternative method for computing the information matrix, for instance Oakes (1999) or Cappé and Moulines (2005).

Let us recall that the scoring function σ is defined, for all $a \in \mathcal{X}$, by $\sigma(a) = \log(q_1(a)/q_0(a))$, we introduce the following function:

$$\begin{aligned} g : \mathbb{R}^{2(d-1)} &\to \mathbb{R}^d \\ \theta &\mapsto \left(\log\left[\frac{e^{\theta_1^1}}{\sum_{k=1}^d e^{\theta_k^1}} \times \frac{\sum_{k=1}^d e^{\theta_k^0}}{e^{\theta_1^0}} \right], \ldots, \log\left[\frac{e^{\theta_d^1}}{\sum_{k=1}^d e^{\theta_k^1}} \times \frac{\sum_{k=1}^d e^{\theta_k^0}}{e^{\theta_d^0}} \right] \right) \end{aligned} \tag{6.20}$$

where $\theta_d^0 = \theta_d^1 = 0$ (by convention) and $J(\theta)$, the Jacobian matrix of g evaluated at $\hat{\theta}$, is detailed thereafter.

$$J(\theta) = \begin{pmatrix} -1+q_0(1) & q_0(2) & \ldots & q_0(d-1) & 1-q_1(1) & -q_1(2) & \ldots & -q_1(d-1) \\ q_0(1) & -1+q_0(2) & \ldots & q_0(d-1) & -q_1(1) & 1-q_1(2) & \ldots & -q_1(d-1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ q_0(1) & q_0(2) & \ldots & -1+q_0(d-1) & -q_1(1) & -q_1(2) & \ldots & 1-q_1(d-1) \\ q_0(1) & q_0(2) & \ldots & q_0(d-1) & -q_1(1) & -q_1(2) & \ldots & -q_1(d-1) \end{pmatrix}.$$

The Jacobian matrix is evaluated at $\hat{\theta}$ and σ is approximated by a first order Taylor expansion of g, i.e. we assumed that $\sigma(.) \sim \mathcal{N}(g(\hat{\theta}), J(\hat{\theta})\hat{\Sigma}J(\hat{\theta})^T)$.

6.2.3 Other types of inference

We propose in this section an overview of other usefulness of the constrained HMM with the computation of a choice of posterior probabilities of interest regarding the existence and the localization of an atypical segment. Let us firstly recall that forward and backward messages are functions of the parameter q. However q is dropped from the notation for the sake of readability.

Posterior location of the atypical segment. For a fixed parameter q, for instance $\hat{q} = \arg\max_q \mathcal{L}(q|S_n \in \{2,3\})$, the posterior probability that the atypical segment starts at position i is given by

$$\mathbb{P}(S_{i-1} = 1, S_i = 2 | X = x, S_n \in \{2,3\}; \hat{q}) = \frac{1}{\mathcal{Z}} F_{i-1}(1)\eta_i(2)B_i(2) \tag{6.21}$$

and the posterior probability that it stops at position i is given by

$$\mathbb{P}(S_i = 2, S_{i+1} = 3 | X = x, S_n \in \{2,3\}; \hat{q}) = \frac{1}{\mathcal{Z}} F_i(2)\eta_i(3)B_{i+1}(3) \tag{6.22}$$

Furthermore, one can compute the maximal scoring segment with the local score using $\hat{\sigma} = \log(\hat{q}_1/\hat{q}_0)$ where $\hat{q}_1$ (respectively $\hat{q}_0$) are the last (respectively first) d values in $\hat{q}$ or alternatively and for the same result, compute $S^* = \text{MAP}(S|X = x; q)$ with the Viterbi (or max-product or max-sum) algorithm over J with the set of potentials $\tilde{\phi}$ in linear in n time complexity to compute the most probable state sequence.

Remark about the transition matrix. The transition function τ simply ensures, with the right initialization of backward messages, a uniform distribution over the chosen state sequence space. Proper transition probabilities are not necessary as we solely need to compute quantities proportional to the likelihood or ratios of quantities with same constant of proportionality. However, proper transition probabilities can be obtained if needed with a backward recursion over the Markov chain $S = (S_1, \ldots, S_n)$, composed solely of hidden states, in order to decompose the state sequence space with no observation and potentials $\phi^0 = \{\tau\}_{i=1,\ldots,n}$. Note that the associated factor graph and junction-tree are also those represented in Figure 6.1.

We denote by $\{B_n^0, \ldots, B_1^0\}$ the set of backward messages recursively computed using a backward pass of the FB algorithm over J with the set of potentials ϕ^0 and initialization $B_n(1) = 0$ and $B_n^0(2) = B_n^0(3) = 1$. For all $i \in \{2, \ldots, n\}$, the probability of transiting from state r at position $i-1$ to state s at position i conditional on $N = 1$ segment is given by

$$\mathbb{P}(S_i = s | S_{i-1} = r, S_n \in \{2, 3\}) = \frac{\tau(r, s) B_i^0(s)}{B_{i-1}^0(r)}$$

Proof. Let us define, for all $i \in \{1, \ldots, n\}$, for $s \in \{1, 2, 3\}$, $F_i^0(s) \propto \mathbb{P}(S_i)$ such that $F_1^0(s) = \tau(1, s)$ and for $i = 2, \ldots, n$, $F_i^0(s) = \sum_{r \in \{1,2,3\}} F_{i-1}(r)\tau(r, s)$, we have $\mathbb{P}(S_i = s | S_{i-1} = r, S_n \in \{2, 3\}) = \frac{\cancel{F_{i-1}^0(r)}\tau(r, s) B_i^0(s)}{\cancel{F_{i-1}^0(r)} B_{i-1}^0(r)}$. □

As an illustrative example, Table 6.1 gives transition probabilities computed for a sequence of length $n = 5$ letters.

Remark: Our particular choice of transition function τ leads to $F_n^0(1) = |\mathcal{S}_n^{\{0\}}| = 1$ and $F_n^0(2) + F_n^0(3) = |\mathcal{S}_n^{\{1\}}| = \binom{n+1}{2}$ where $\mathcal{S}_n^{\{0\}}$ (respectively $\mathcal{S}_n^{\{1\}}$) is the set of sequences containing exactly zero (respectively one) atypical segment.

Existence of an atypical segment We have seen in the previous paragraph that a direct consequence of our assumption of a uniform distribution over the state sequence space ensured by τ implies that $\mathbb{P}(S_n = 1) = \mathbb{P}(N = 0) \propto F_n^0(1)$ and $\mathbb{P}(S_n \in \{2, 3\}) = \mathbb{P}(N = 1) \propto F_n^0(2) + F_n^0(3)$ where $F_n^0(1) = |\mathcal{S}_n^0| = 1$ and $F_n^0(2) + F_n^0(3) = |\mathcal{S}_n^1| = \binom{n+1}{2}$. In order to avoid such a strong prior of the number of segments, let us define

$$\mathbb{Q}(N, X = x|\hat{q}) \stackrel{\text{def}}{=} \mathbb{P}(X = x|N; \hat{q})\mathbb{Q}(N)$$

	r=1	r=2	r=3
s=1	0.60	0.40	0.00
s=2	0.00	0.80	0.20
s=3	0.00	0.00	1.00

(a) $\mathbb{P}(S_2 = s|S_1 = r)$

	r=1	r=2	r=3
s=1	0.50	0.50	0.00
s=2	0.00	0.75	0.25
s=3	0.00	0.00	1.00

(b) $\mathbb{P}(S_3 = s|S_2 = r)$

	r=1	r=2	r=3
s=1	0.33	0.67	0.00
s=2	0.00	0.67	0.33
s=3	0.00	0.00	1.00

(c) $\mathbb{P}(S_4 = s|S_3 = r)$

	r=1	r=2	r=3
s=1	0.00	1.00	0.00
s=2	0.0	0.50	0.50
s=3	0.0	0.00	1.00

(d) $\mathbb{P}(S_5 = s|S_4 = r)$

Table 6.1: Transition probabilities computed for a sequence of length $n = 5$.

and therefore

$$\mathbb{Q}(N|X = x; \hat{q}) = \frac{\mathbb{P}(X = x|N; \hat{q})\mathbb{Q}(N)}{\mathbb{Q}(X = x|\hat{q})} \propto \mathbb{P}(X = x|N; \hat{q})\mathbb{Q}(N)$$

where $\mathbb{Q}(N)$ is an arbitrary prior for $N \in \{0, 1\}$. Note that we have

$$\mathbb{P}(X = x|N = 1; \hat{q}) = \mathbb{P}(X = x|S_n \in \{2, 3\}; \hat{q}) = \frac{F_n(2) + F_n(3)}{F_n^0(2) + F_n^0(3)}$$

and therefore, the posterior probability that the sequence contains an atypical segment is given by

$$\mathbb{Q}(N = 1|X = x; \hat{q}) \propto \frac{F_n(2) + F_n(3)}{F_n^0(2) + F_n^0(3)}\mathbb{Q}(N = 1). \tag{6.23}$$

An extension of the state space of the number of segments with an arbitrary prior distribution constitute the main goal of Section 6.3.

6.2.4 Applications

In this section, we provide two application examples, one over simulated datasets and one over real datasets of transmembrane (TM) proteins.

6.2.4.1 Simulated datasets

Estimates of Parameter and Confidence Intervals (CIs). In the first simulation scheme, we simulated 50, on one hand, and 500, on the other hand, sequences over the alphabet $\mathcal{X} = \{1, \ldots, d = 7\}$. Sequence lengths are randomly simulated between $n = 150$ and $n = 1000$ letters. Each sequence contains one atypical segment of random length at a random location in the sequence. Simulations have been done with true parameter θ^* given in Table 6.2. Because that simulation scheme will be later reused in this section, we give it the names *simu1* for simplicity. Estimates for θ and σ and their 95% CIs estimated with the observed Fisher information are

	θ^*	$\hat{\theta}$	[95% $\widehat{\text{CI}}$], 50 seq.	$\hat{\theta}$	[95% $\widehat{\text{CI}}$], 500 seq.
θ_1^0	0.0000	0.0075	[−0.0461; 0.0612]	0.0010	[−0.0153; 0.0173]
θ_2^0	0.0000	0.0493	[−0.0028; 0.1013]	0.0028	[−0.0132; 0.0189]
θ_3^0	0.0000	−0.0186	[−0.0716; 0.0343]	−0.0026	[−0.0187; 0.0135]
θ_4^0	0.0000	−0.0077	[−0.0609; 0.0454]	−0.0008	[−0.0169; 0.0154]
θ_5^0	0.0000	0.0072	[−0.0457; 0.0602]	−0.0042	[−0.0204; 0.0119]
θ_6^0	0.0000	−0.0079	[−0.0608; 0.0450]	−0.0101	[−0.0262; 0.0061]
θ_1^1	1.3710	1.3811	[1.2897; 1.4725]	1.3518	[1.3212; 1.3825]
θ_2^1	−0.5647	−0.4544	[−0.5815; −0.3274]	−0.5729	[−0.6181; −0.5276]
θ_3^1	0.3631	0.3680	[0.2650; 0.4711]	0.3466	[0.3116; 0.3815]
θ_4^1	0.6329	0.6200	[0.5208; 0.7192]	0.6258	[0.5925; 0.6592]
θ_5^1	0.4043	0.3936	[0.2912; 0.4961]	0.3891	[0.3545; 0.4238]
θ_6^1	−0.1061	−0.1524	[−0.2700; −0.0349]	−0.1455	[−0.1848; −0.1061]

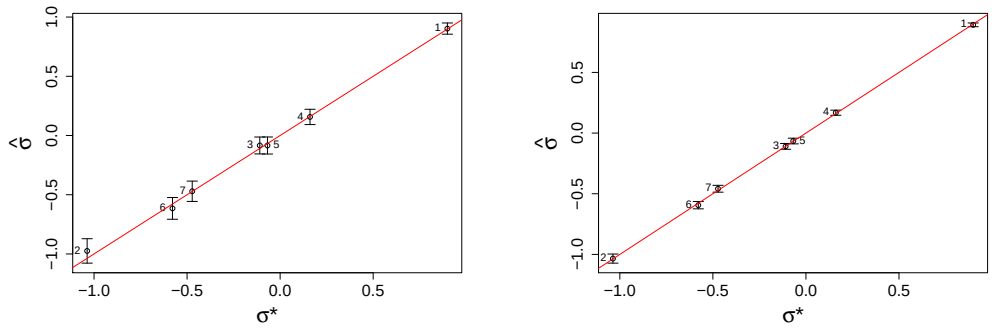
Table 6.2: Estimates of θ and estimates of 95% CIs computed with the set of 50 simulated sequences (left) and 500 simulated sequences (right) according to the simulation scheme simu1. The true parameter used for the simulations is denoted θ^*.

respectively given in Table 6.2 and Table 6.3. A graphical representation of Table 6.3 is proposed Figure 6.3. We notice that, as expected, all CIs shrink of about a third with ten times more simulations ($1/\sqrt{10} \approx 1/3$). Furthermore, for both simulated datasets, true parameters θ^* and σ^* fall inside the estimated confidence interval 95% $\widehat{\text{CI}}$.

Coverage probabilities have been empirically studied in other simulation schemes. Table 6.4 gives the estimated (empirical mean) coverage probability of 95 %, 90% and 70 % estimated $\widehat{\text{CI}}$ for θ and σ as well as their own 95% CIs in two different simulation schemes. In both schemes we simulated 200 sets of 150 sequences over $\mathcal{X} = \{1, 2, 3, 4\}$ with true parameter $\theta^* = (0.0000, 0.0000, 0.0000, 1.3710, -0.5647, 0.3631)$. Each se-

	σ^*	$\hat{\sigma}$	[95% $\widehat{\text{CI}}$], 50 seq.	$\hat{\sigma}$	[95% $\widehat{\text{CI}}$], 500 seq.
$\sigma(1)$	0.8986	0.9028	[0.8554; 0.9501]	0.8921	[0.8767; 0.9075]
$\sigma(2)$	−1.0370	−0.9745	[−1.0776; −0.8713]	−1.0344	[−1.0719; −0.9969]
$\sigma(3)$	−0.1092	−0.0841	[−0.1554; −0.0128]	−0.1095	[−0.1335; −0.0856]
$\sigma(4)$	0.1605	0.1569	[0.0924; 0.2215]	0.1679	[0.1468; 0.1890]
$\sigma(5)$	−0.0681	−0.0844	[−0.1564; −0.0124]	−0.0654	[−0.0889; −0.0419]
$\sigma(6)$	−0.5784	−0.6153	[−0.7070; −0.5236]	−0.5941	[−0.6243; −0.5639]
$\sigma(7)$	−0.4723	−0.4708	[−0.5565; −0.3851]	−0.4587	[−0.4870; −0.4305]

Table 6.3: Estimates of the SF σ and estimates of 95% CIs computed with the set of 50 simulated sequences (left) and 500 simulated sequences (right) according to the simulation scheme simu1. The true scoring function is denoted by σ^*.



(a) 50 simulated sequences

(b) 500 simulated sequences

Figure 6.3: Graphical representation of Table 6.3. Estimates $\hat{\sigma}$ of the SF σ with 95% CIs versus true values σ^* computed over the 50 simulated sequences (on the left) and the 500 simulated sequences (on the right) according to simulation scheme simu1.

quence contains an atypical segment of length 40 letters at a random position. Simulated sequences are of length $n = 150$ (respectively $n = 1000$) letters in the first (respectively the second) simulation scheme. We used the EM algorithm for the estimation of θ with five start points and kept the estimate $\hat{\theta}$ associated with the highest log-likelihood. As shown in Table 6.4a, coverage probabilities of CIs for θ are the expected ones whereas estimates are slightly biased for σ. This biais is explained by the approximation made with a non linear transformation of the multivariate Gaussian distribution g defined in Equation (6.20) using a first order Taylor expansion. Furthermore, comparing Tables 6.4a and 6.4b and recalling that all atypical segments are of length 40 letters, one can see that the longer the sequences are, the more biased estimates are. Indeed, this is the main limitation of the method which requires a sufficient length ratio between atypical segment and non-atypical segments. A study is still in preparation for the evaluation of the asymptotic behavior of the estimator in various simulation schemes especially various length ratios, sequence length, parameter and parameter dimension (i.e. cardinal of the alphabet).

Posterior localization of the atypical segment. Secondly we computed posterior probabilities of interest such as marginal posterior state probabilities (Figure 6.4a), posterior probabilities of start and stop positions of the atypical segment (Figures 6.4b) and the most probable location of the atypical segment in two sequences computed according to simulation scheme simu1. Posterior probabilities are computed with estimate $\hat{q} = (\hat{q}_0, \hat{q}_1)$ obtained from $\hat{\theta}$ in Table 6.2. The true position of the atypical segment is delimited by the greyed region (from index 356 to index 601 for the first sequence and from 433 to 474 for the second one). The maximum scoring segment computed with $\hat{\sigma} = \log(\hat{q}_1/\hat{q}_0)$ (or equivalently the most probable sequence state computed with the max-sum algorithm and $\hat{q}$) returned positions 353 to 599 (respectively 414 to 476) for the first (respectively second) sequence. Let us

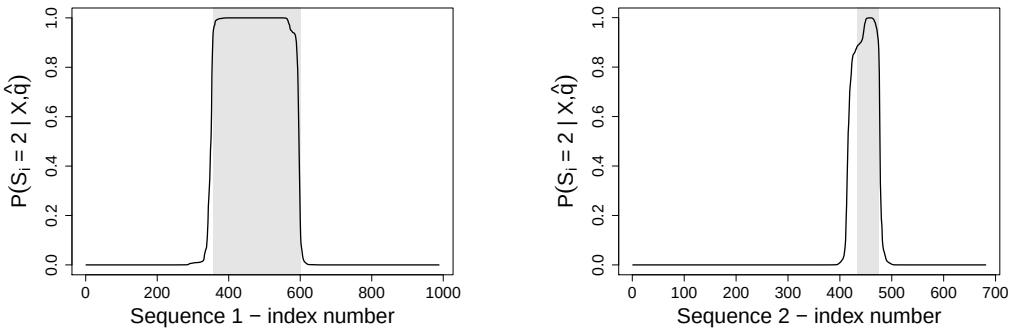
Sequence length 150	95% CI	90% CI	70% CI
θ_1^0	0.930 [0.895; 0.965]	0.890 [0.845; 0.930]	0.700 [0.635; 0.760]
θ_2^0	0.950 [0.920; 0.980]	0.895 [0.850; 0.935]	0.745 [0.685; 0.805]
θ_3^0	0.955 [0.925; 0.980]	0.890 [0.845; 0.930]	0.705 [0.640; 0.765]
θ_1^1	0.930 [0.895; 0.965]	0.890 [0.845; 0.930]	0.715 [0.650; 0.775]
θ_2^1	0.970 [0.945; 0.990]	0.910 [0.870; 0.950]	0.735 [0.675; 0.795]
θ_3^1	0.965 [0.940; 0.990]	0.910 [0.870; 0.950]	0.750 [0.690; 0.810]
$\sigma(1)$	0.935 [0.900; 0.965]	0.880 [0.835; 0.925]	0.695 [0.630; 0.760]
$\sigma(2)$	0.915 [0.875; 0.950]	0.830 [0.775; 0.880]	0.625 [0.555; 0.690]
$\sigma(3)$	0.925 [0.885; 0.960]	0.895 [0.850; 0.935]	0.685 [0.620; 0.750]
$\sigma(4)$	0.950 [0.920; 0.980]	0.910 [0.870; 0.950]	0.690 [0.625; 0.755]

(a) Sequences of length 150 letters

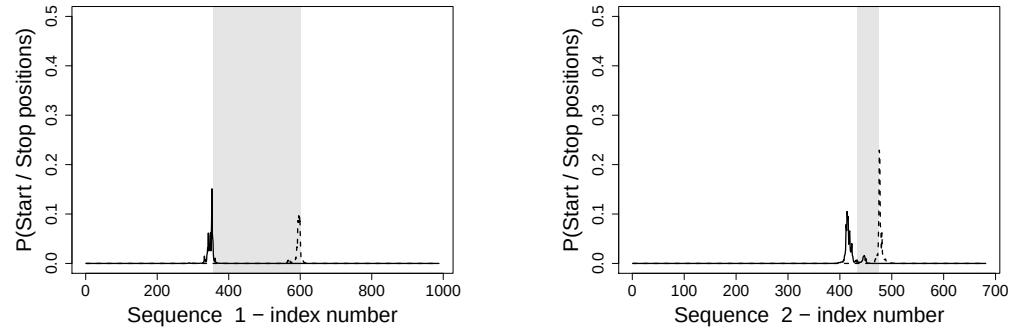
Sequence length 1000	95% CI	90% CI	70% CI
θ_1^0	0.985 [0.965; 1.000]	0.920 [0.880; 0.955]	0.700 [0.635; 0.760]
θ_2^0	0.970 [0.945; 0.990]	0.950 [0.920; 0.980]	0.710 [0.645; 0.770]
θ_3^0	0.935 [0.900; 0.965]	0.920 [0.880; 0.955]	0.750 [0.690; 0.810]
θ_1^1	0.950 [0.920; 0.980]	0.895 [0.850; 0.935]	0.695 [0.630; 0.760]
θ_2^1	0.925 [0.885; 0.960]	0.890 [0.845; 0.930]	0.695 [0.630; 0.760]
θ_3^1	0.955 [0.925; 0.980]	0.905 [0.865; 0.945]	0.690 [0.625; 0.755]
$\sigma(1)$	0.935 [0.900; 0.965]	0.895 [0.850; 0.935]	0.660 [0.595; 0.725]
$\sigma(2)$	0.930 [0.895; 0.965]	0.850 [0.800; 0.900]	0.625 [0.555; 0.690]
$\sigma(3)$	0.945 [0.910; 0.975]	0.880 [0.835; 0.925]	0.680 [0.615; 0.745]
$\sigma(4)$	0.960 [0.930; 0.985]	0.925 [0.885; 0.960]	0.700 [0.635; 0.760]

(b) Sequences of length 1000 letters

Table 6.4: Coverage probability of 95% CI, 90% CI and 70% CI for θ and σ estimated with the observed Fisher information over 200 sets of 150 sequences of length 150 letters (Table 6.4a) or 1000 letters (Table 6.4b) and their own 95% CI. All sequences contain an atypical segment of length 40 letters at a random position. Greyed lines highlight a situation where the corresponding 95% CI for the coverage probability does not contain the expected coverage probability

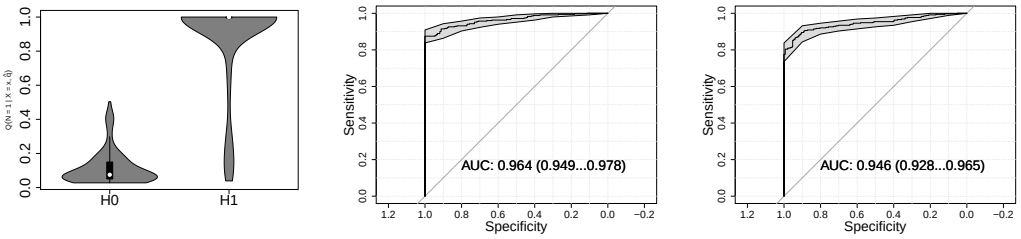


(a) Marginal posterior probability of each state to be atypical in the first sequence (left) and the second sequence (right)



(b) Posterior probabilities of start (plain line) and stop (dashed line) position of the atypical segment for the first sequence (left) and the second sequence (right)

Figure 6.4: Marginal posterior probability of each state to be atypical computed with $\hat{q}$ (Figure 6.4a) and marginal posterior probability of the start (plain lines) and stop (dashed lines) positions of the segment (Figures 6.4b) for the first two sequences containing an atypical segment in the dataset of 500 simulated sequences. The true position of the segment is delimited by the greyed region.



(a) Violin plot for $\mathbb{Q}(N = 1|\hat{q})$ (b) AUROC associated with Bayes factor BF. (c) AUROC associated with a Gumbel approximation.

Figure 6.5: Violin plot for posterior probabilities that sequences contain an atypical segment using estimate $\hat{q}$, AUROC of the Bayes factor $\mathbb{Q}_{H_0}(N = 1|X = x, \hat{q})/\mathbb{Q}_{H_1}(N = 1|X = x, \hat{q})$ and AUROC of a Gumbel approximation.

define a true positive (TP) (respectively a false positive (FP)) to be a letter inside the atypical segment detected as such (respectively detected as outside the atypical segment) by the most probable state sequence or the maximal scoring segment, and a true negative (TN) (respectively a false negative (FN)) to be a letter outside the atypical segment detected as such (respectively detected as inside the atypical segment). The sensitivity (TP / (TP+FN)) and the specificity (TN / (TN+FP)) of the maximal scoring segment computed with $\hat{\sigma}$ over the whole dataset of 500 sequences are respectively 0.9551 and 0.9787.

Posterior probability of the existence of an atypical segment. Thirdly we simulated 500 sequences of random length between 150 and 1000 letters over the alphabet $\mathcal{X} = \{1, \ldots, d = 7\}$, 70 % of which according to simulation scheme simu1 contain one atypical segment of random length and random location and 30% of which contain no atypical segment. Let us propose a hypothesis testing $T^{\{0,1\}}$ in order to question whether a sequence contains an atypical segment (H_0) or not (H_1) based on the Bayes factor

$$\text{BF} = \frac{\mathbb{Q}_{H_0}(N = 1|X = x, \hat{q})}{\mathbb{Q}_{H_1}(N = 1|X = x, \hat{q})}$$

computed with Equation (6.23) where we assume that the prior $\mathbb{Q}(N)$ is uniform over $\{0, 1\}$ and $\mathbb{Q}_{H_0}$ (respectively $\mathbb{Q}_{H_1}$) denotes the probability distribution $\mathbb{Q}$ conditional on H_0 (respectively H_1). Figure 6.5 is a graphical representation of different tests regarding the posterior probabilities that sequences contain an atypical segment with violin plots of $\mathbb{Q}_{H_1}(N = 1|X = x; \hat{q})$ and $\mathbb{Q}_{H_0}(N = 1|X = x; \hat{q})$ (Figure 6.5a), an AUROC curve of BF leading to an AUC equal to 0.9638 with 95% CI: $[0.9495; 0.9781]$, DeLong (Figure 6.5b) and and AUROC curve of p-values computed with a Gumbel approximation whose parameters were fitted empirically on a sample of size 1000 which led to an AUC equal to 0.9462 with 95% CI: $[0.9275; 0.9648]$, DeLong (Figure 6.5c). These results suggest a very high accuracy of $T^{\{0,1\}}$ using $\hat{q}$ in selecting sequences containing an atypical segment out of a set of mixed sequences.

Remark. This whole application section still needs to be densified with a variety of simulation schemes in order to evaluate the properties of the estimator, sensitivity and specificity of the maximal scoring segment using $\hat{\sigma}$ as well as performances of $T^{\{0,1\}}$ in various simulation schemes, especially various sequence lengths, length ratios between atypical and non-atypical segments, parameter values and dimension (i.e. cardinal of the alphabet).

6.2.4.2 Real dataset of transmembrane proteins

We consider in this section the real dataset of single-pass TM proteins extracted from the UniProt database[2] with the request "single-pass membrane protein type 1" and filters "reviewed:yes" and "organism:Homo sapiens (Human)". We left two of those sequences, randomly chosen, appart: O43493 (Trans-Golgi network integral membrane protein 2) and Q9H3N1 (Thioredoxin-related transmembrane protein 1). Amino acids are labelled with the single letter amino acid code of the International Nucleotide Sequence Database[3], hence $\mathcal{X} = \{\mathtt{A}, \mathtt{R}, \mathtt{D}, \mathtt{N}, \mathtt{C}, \mathtt{E}, \mathtt{Q}, \mathtt{G}, \mathtt{H}, \mathtt{I}, \mathtt{L}, \mathtt{K}, \mathtt{M}, \mathtt{F}, \mathtt{P}, \mathtt{S}, \mathtt{T}, \mathtt{W}, \mathtt{Y}, \mathtt{V}\}$. In this section, a hat symbol denotes an estimate computed with our method and a tilda symbol denotes an estimate computed with the known true proportion of amino acids in TM and non-TM segments obtained from the `.gff` file downloaded from the UniProt website which indicates the known location of TM segments.

Estimation. EM was initialized with proportions of letters in the whole dataset for q_0 and with a random distribution for q_1. We used 25 random start points for q_1, keeping the estimate associated with the highest log-likelihood. Due to the sensitivity of the method to the initialization of the EM algorithm, we solely kept sequences of length at most 500 amino acids in order to ensure a sufficient length ratio between TM and non-TM segments. The method applied to the complete dataset (with no restriction on sequence length) led to unstable estimates if EM was initialized with a random parameter q_1. However estimates where close to those computed over the restricted dataset when initializing either with the Kyte and Doolittle scale (Kyte and Doolittle, 1982) or the TM tendency scale (Zhao and London, 2006) rescaled with the right temperature to ensure the equivalence between the Gibbs measure and the generative model (see Theorems 5 and 6).

Estimates of θ (respectively σ) and of 95% CIs are given in Table 6.5 and 6.6 (respectively Table 6.7) with $\hat{\theta}^0_{\mathtt{V}} = \hat{\theta}^1_{\mathtt{V}} = \tilde{\theta}^0_{\mathtt{V}} = \tilde{\theta}^1_{\mathtt{V}} = 0$. A graphical representation of Table 6.7 is proposed in Figure 6.6. As expected, the rarer amino acids in TM segments are (e.g. $\mathtt{R}$, $\mathtt{N}$, $\mathtt{E}$, $\mathtt{H}$ versus $\mathtt{I}$, $\mathtt{V}$, $\mathtt{L}$, $\mathtt{F}$), the wider the CIs for estimates of θ^1 and σ are. The parameter $\tilde{\theta}^0$ falls inside the corresponding $\widehat{\mathrm{CI}}$ whereas biased estimates for θ^1 and σ highlight other weaknesses of the method in difficult frameworks. In this context firstly the parameter is of high (18) dimension. Secondly, the extreme rarity of some amino acids in TM segments lead to unstable estimates. One solution to overcome that problem could be to fix a very low frequency for those extremely rare amino acids and estimate solely parameters for more frequent ones. Note that

[2] `http://www.uniprot.org/uniprot/`

[3] `http://www.insdc.org`

	$\tilde{\theta}^0$	95% $\widetilde{\text{CI}}$	$\hat{\theta}^0$	95% $\widehat{\text{CI}}$
$\theta^0_{\mathtt{A}}$	0.0922	[0.0614; 0.1230]	0.0784	[0.0473; 0.1095]
$\theta^0_{\mathtt{R}}$	−0.1078	[−0.1401; −0.0754]	−0.1174	[−0.1499; −0.0849]
$\theta^0_{\mathtt{D}}$	−0.3231	[−0.3575; −0.2888]	−0.3352	[−0.3697; −0.3007]
$\theta^0_{\mathtt{N}}$	−0.4884	[−0.5246; −0.4523]	−0.4994	[−0.5356; −0.4632]
$\theta^0_{\mathtt{C}}$	−0.7903	[−0.8302; −0.7504]	−0.8056	[−0.8460; −0.7652]
$\theta^0_{\mathtt{E}}$	−0.0083	[−0.0399; 0.0233]	−0.0192	[−0.0508; 0.0125]
$\theta^0_{\mathtt{Q}}$	−0.3226	[−0.3570; −0.2882]	−0.3326	[−0.3671; −0.2981]
$\theta^0_{\mathtt{G}}$	0.0994	[0.0686; 0.1301]	0.0875	[0.0566; 0.1185]
$\theta^0_{\mathtt{H}}$	−0.8858	[−0.9270; −0.8446]	−0.892	[−0.9334; −0.8507]
$\theta^0_{\mathtt{I}}$	−0.6032	[−0.6406; −0.5657]	−0.5897	[−0.6273; −0.5521]
$\theta^0_{\mathtt{L}}$	0.4369	[0.4083; 0.4654]	0.4313	[0.4023; 0.4602]
$\theta^0_{\mathtt{K}}$	−0.3451	[−0.3797; −0.3105]	−0.3573	[−0.3920; −0.3225]
$\theta^0_{\mathtt{M}}$	−1.2248	[−1.2715; −1.1781]	−1.2237	[−1.2707; −1.1766]
$\theta^0_{\mathtt{F}}$	−0.6990	[−0.7376; −0.6603]	−0.6859	[−0.7246; −0.6472]
$\theta^0_{\mathtt{P}}$	0.1779	[0.1478; 0.2081]	0.1686	[0.1383; 0.1990]
$\theta^0_{\mathtt{S}}$	0.3604	[0.3314; 0.3894]	0.3506	[0.3215; 0.3798]
$\theta^0_{\mathtt{T}}$	0.0657	[0.0347; 0.0967]	0.0568	[0.0257; 0.0880]
$\theta^0_{\mathtt{W}}$	−1.3427	[−1.3917; −1.2938]	−1.3179	[−1.3670; −1.2688]
$\theta^0_{\mathtt{Y}}$	−0.8119	[−0.8521; −0.7717]	−0.8058	[−0.8461; −0.7655]

Table 6.5: Estimates of θ^0 and of 95% CIs computed from the real dataset of TM proteins respectively with the unsupervised method ($\hat{\theta}^0$ and $\widehat{\text{CI}}$) and with the true proportion of amino acids in the dataset ($\tilde{\theta}^0$ and $\widetilde{\text{CI}}$).

we obtained a good correlation between our estimate $\hat{\sigma}$ and respectively the Kyte and Doolittle scale (Figures 6.7a, correlation = 0.92) and the TM tendency scale (Figure 6.7b, correlation = 0.87).

Posterior localization of TM segment. A graphical representation of posterior probabilities of interest computed with $\hat{q}$ for the two sequences O43493 and Q9H3N1 is proposed Figure 6.8 where the greyed region delimits the true location of the TM segment (382-402 for sequence O43493 and 181-203 for sequence Q9H3N1). Figure 6.8a (respectively Figure 6.8b) represent the marginal posterior state probabilities (respectively the posterior probabilities of start and stop positions of the TM segment). Furthermore, Table 6.8 gives the maximal scoring segment (i.e. the most probable start and stop positions of the TM-segment) computed with three different SFs: the unsupervised estimate $\hat{\sigma}$, the Kyte and Doolittle scale (KD scale) and Zhao's TM tendency scale (TM scale).

The extension to the analysis of the sensitivity and specificity of the maximal scoring segment computed with four different SFs over three datasets is detailed in Table 6.9. The four SFs used are the unsupervised estimate $\hat{\sigma}$, estimates $\tilde{\sigma}_D$ computed with the known proportion of letters in the studied dataset D, the KD scale

	$\tilde{\theta}^1$	95% $\widetilde{\text{CI}}$	$\hat{\theta}^1$	95% $\widehat{\text{CI}}$
$\theta^1_{\texttt{A}}$	-0.4155	$[-0.4957; -0.3353]$	-0.3409	$[-0.4303; -0.2515]$
$\theta^1_{\texttt{R}}$	-3.8495	$[-4.1996; -3.4993]$	-5.8641	$[-8.8088; -2.9194]$
$\theta^1_{\texttt{D}}$	-4.5426	$[-5.0352; -4.0500]$	-5.2772	$[-6.2367; -4.3178]$
$\theta^1_{\texttt{N}}$	-4.4820	$[-4.9600; -4.0040]$	-5.8022	$[-7.4664; -4.1380]$
$\theta^1_{\texttt{C}}$	-1.7623	$[-1.8943; -1.6302]$	-1.6667	$[-1.8201; -1.5133]$
$\theta^1_{\texttt{E}}$	-4.0194	$[-4.3999; -3.6388]$	-5.5313	$[-6.8295; -4.2330]$
$\theta^1_{\texttt{Q}}$	-3.9140	$[-4.2754; -3.5526]$	-4.9748	$[-5.8621; -4.0875]$
$\theta^1_{\texttt{G}}$	-0.6885	$[-0.7759; -0.6011]$	-0.6356	$[-0.7324; -0.5388]$
$\theta^1_{\texttt{H}}$	-3.8812	$[-4.2369; -3.5256]$	-5.2708	$[-6.8081; -3.7336]$
$\theta^1_{\texttt{I}}$	-0.1708	$[-0.2456; -0.0961]$	-0.1939	$[-0.2769; -0.1108]$
$\theta^1_{\texttt{L}}$	0.3990	$[\ 0.3337;\ 0.4644]$	0.4283	$[\ 0.3545;\ 0.5021]$
$\theta^1_{\texttt{K}}$	-3.6016	$[-3.9119; -3.2914]$	-3.7527	$[-4.2309; -3.2745]$
$\theta^1_{\texttt{M}}$	-1.9307	$[-2.0728; -1.7887]$	-2.0264	$[-2.2037; -1.8490]$
$\theta^1_{\texttt{F}}$	-0.8398	$[-0.9318; -0.7477]$	-0.9416	$[-1.0494; -0.8338]$
$\theta^1_{\texttt{P}}$	-2.1853	$[-2.3443; -2.0263]$	-2.3816	$[-2.6190; -2.1442]$
$\theta^1_{\texttt{S}}$	-1.3213	$[-1.4314; -1.2111]$	-1.3606	$[-1.4940; -1.2272]$
$\theta^1_{\texttt{T}}$	-1.4017	$[-1.5155; -1.2880]$	-1.4407	$[-1.5751; -1.3064]$
$\theta^1_{\texttt{W}}$	-2.1794	$[-2.3380; -2.0208]$	-2.7335	$[-3.0361; -2.4309]$
$\theta^1_{\texttt{Y}}$	-2.1112	$[-2.2650; -1.9574]$	-2.5101	$[-2.7667; -2.2535]$

Table 6.6: Estimates of θ^1 and of 95% CIs computed from the real dataset of TM proteins respectively with the unsupervised method ($\hat{\theta}^1$ and $\widehat{\text{CI}}$) and with the true proportion of amino acids in the dataset ($\tilde{\theta}^1$ and $\widetilde{\text{CI}}$). A greyed line signifies that the parameter $\tilde{\theta}^1_a$ falls outside the corresponding $\widetilde{\text{CI}}$.

	$\tilde{\sigma}$	[95% $\widetilde{\text{CI}}$]	$\hat{\sigma}$	[95% $\widehat{\text{CI}}$]
σ(A)	0.4359	[0.3736; 0.4982]	0.5404	[0.4703; 0.6106]
σ(R)	−2.7981	[−3.1447; −2.4514]	−4.7870	[−7.7289; −1.8451]
σ(D)	−3.2759	[−3.7661; −2.7856]	−3.9823	[−4.9398; −3.0248]
σ(N)	−3.0499	[−3.5257; −2.5742]	−4.3431	[−6.0067; −2.6794]
σ(C)	−0.0283	[−0.1530; 0.0964]	0.0986	[−0.0477; 0.2449]
σ(E)	−3.0674	[−3.4447; −2.6902]	−4.5524	[−5.8503; −3.2545]
σ(Q)	−2.6478	[−3.0060; −2.2896]	−3.6825	[−4.5676; −2.7973]
σ(G)	0.1558	[0.0844; 0.2272]	0.2366	[0.1576; 0.3156]
σ(H)	−2.0518	[−2.4049; −1.6987]	−3.4191	[−4.9550; −1.8831]
σ(I)	1.3760	[1.3169; 1.4351]	1.3555	[1.2883; 1.4228]
σ(L)	0.9058	[0.8659; 0.9457]	0.9568	[0.9109; 1.0027]
σ(K)	−2.3129	[−2.6194; −2.0064]	−2.4357	[−2.9100; −1.9614]
σ(M)	0.2377	[0.1004; 0.3751]	0.1570	[−0.0185; 0.3325]
σ(F)	0.8028	[0.7224; 0.8833]	0.7040	[0.6074; 0.8006]
σ(P)	−1.4196	[−1.5703; −1.2689]	−1.5905	[−1.8190; −1.3620]
σ(S)	−0.7380	[−0.8354; −0.6407]	−0.7515	[−0.8720; −0.6311]
σ(T)	−0.5238	[−0.6258; −0.4218]	−0.5379	[−0.6596; −0.4161]
σ(W)	0.1069	[−0.0482; 0.2620]	−0.4559	[−0.7588; −0.1529]
σ(Y)	−0.3557	[−0.5033; −0.2081]	−0.7446	[−0.9978; −0.4913]
σ(V)	0.9436	[0.8925; 0.9947]	0.9597	[0.9023; 1.0171]

Table 6.7: Estimates of the SF σ and 95% CIs computed over the real dataset of TM proteins. A greyed line signifies that the parameter $\tilde{\sigma}(a)$ falls outside the corresponding $\widehat{\text{CI}}$.

		true	$\hat{\sigma}$	KD scale	TM scale
seq. O43493	start	382	385	3	385
	stop	402	402	22	402
seq. Q9H3N1	start	181	183	138	138
	stop	203	203	207	203

Table 6.8: Most probable start and stop position of the TM segment computed with the maximal scoring segment using SFs $\hat{\sigma}$ the Kyte and Doolittle scale (KD scale) and the TM tendency scale (TM scale).

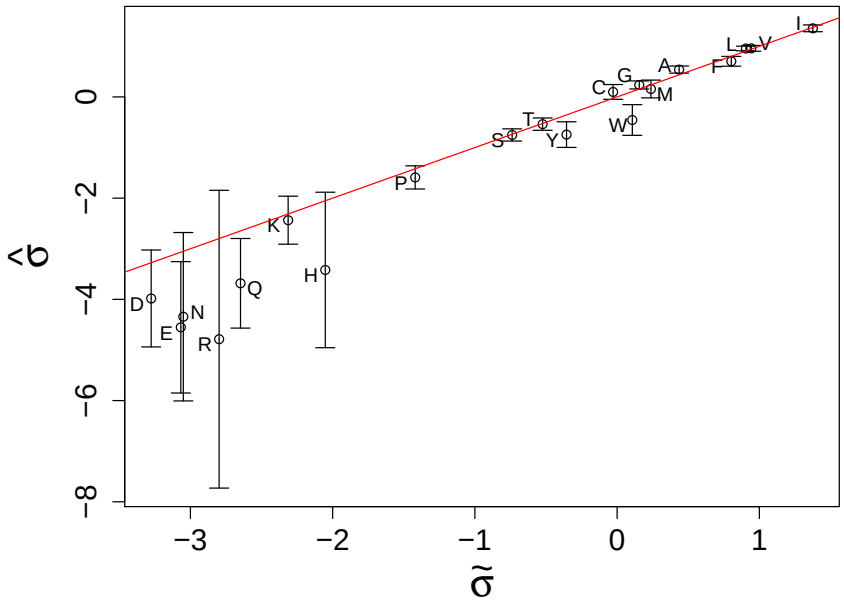
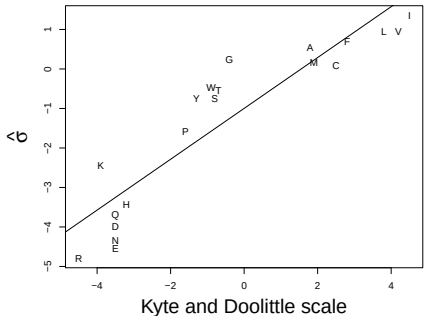
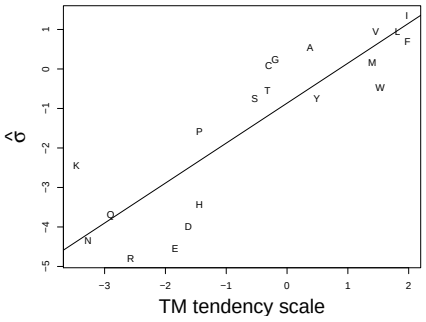


Figure 6.6: Linear regression of $\hat{\sigma}$ on $\tilde{\sigma}$ computed from estimates of θ reported in Tables 6.5 and 6.6 along with 95% CI for $\hat{\sigma}$ (correlation = 0.9998794).

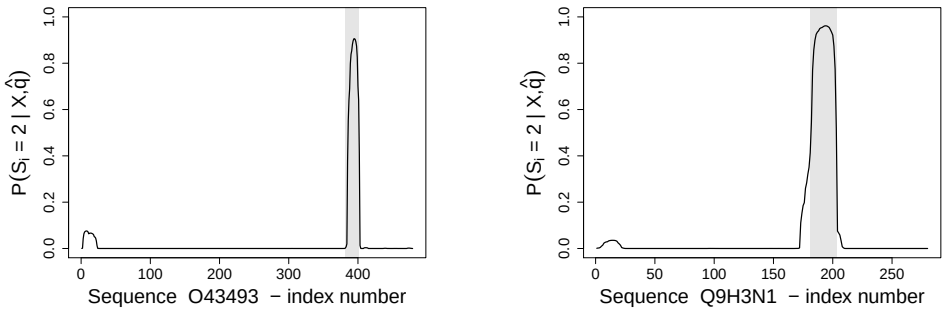


(a) Linear regression of the estimate $\hat{\sigma}$ on the Kyte and Doolittle scale (correlation = 0.92).

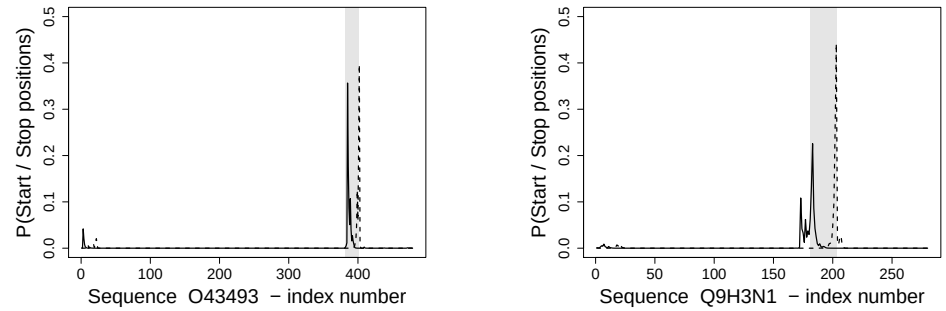


(b) Linear regression of the estimate $\hat{\sigma}$ on the TM tendency scale (correlation = 0.87).

Figure 6.7: Linear regression of the estimate $\hat{\sigma}$ on the Kyte and Doolittle scale and on the TM tendency scale.



(a) Marginal posterior probability of each state to be atypical for sequence O43493 (left) and sequence Q9H3N1 (right).



(b) Posterior probabilities of start (plain line) and stop (dashed line) position of the atypical segment for sequence O43493 (left) and sequence Q9H3N1 (right).

Figure 6.8: Marginal posterior probability of each state to be atypical computed with $\hat{q}$ (Figure 6.8a) and posterior probabilities of start (plain lines) and stop (dashed lines) positions of the TM segment (Figures 6.8b) for the two sequences left appart (O43493 and Q9H3N1). The true position of the TM segment is delimited by the two dashed lines.

and the TM scale. The first dataset is the one used to compute $\hat{\sigma}$, i.e. the dataset restricted to sequences no longer than 500 amino acids, the second dataset is the whole dataset including sequences of length greater than 500 amino acids, the third dataset is a dataset of $12,733$ proteins uploaded from the UniProt website with no restriction on organisms (downloaded with the request "single-pass membrane protein" AND Reviewed = yes). Results shown in Table 6.9 suggest a better specificity of the local score when using $\hat{\sigma}$ rather than the Kyte and Doolittle scale and the TM tendency scale but a lower sensitivity which even drops to 0.76 with the new dataset. These results suggest that the maximal scoring segment computed with $\hat{\sigma}$ misses suboptimal segments in comparison with commonly used TM scales. This could be the consequence of a unique temperature T rendering both models defined Equation 6.13 and 6.14 equivalent using SF $\sigma = f/T$ with $f = T \log(q_1/q_0)$ (see Theorems 5 and 6). Indeed Karlin et al. explained in (Karlin and Altschul, 1990) and (Karlin, 2005) that the function $\sigma = \log(q_1/q_0)$ always satisfies Assumptions 6.11 and 6.12 is $q_1 \neq q_0$ but combining them with experimental data could better highlight the feature of interest. However, the very high specificity of our method and its unsupervised nature could be of interest in pointing at segments of unknown feature. Note that the specificity of our method applied to the real datasets overpasses the one applied to simulated datasets suggesting again a high dependence on several parameters such as segment / sequence length ratios as well as q_1/q_0 ratios. Hence, we still have to perform further and more convincing investigations both on simulated datasets with a variety of simulation schemes as well as apply our method to other real datasets for which our method would have a significant impact. In particular, its unsupervised nature could be of important usefulness if supervised and empirical SF are difficult to establish, in particular if the feature of interest is insufficiently known. An article is still in preparation with these complementary studies.

		$\hat{\sigma}$	$\tilde{\sigma}_D$	KD scale	TM scale
Dataset 1	Sensitivity	0.8871	0.9030	0.9125	0.9419
	Specificity	0.9961	0.9958	0.9316	0.9900
Dataset 2	Sensitivity	0.8492	0.8791	0.9121	0.9231
	Specificity	0.9905	0.9897	0.8800	0.9760
Dataset 3	Sensitivity	0.7592	0.8774	0.9064	0.9263
	Specificity	0.9930	0.9873	0.8827	0.9726

Table 6.9: Sensitivity and specificity of the maximal scoring segment computed with four different SFs: the unsupervised estimate $\hat{\sigma}$, the estimate $\tilde{\sigma}_D$ computed with the true proportion of amino acids in TM and non-TM segments in the studied datatset D, the Kyte and Doolittle scale (KD scale) and Zhao's TM tendency scale (TM scale). The first, second and third datasets are respectively the one used to compute $\hat{\sigma}$, the whole dataset with no restriction on sequence length and a dataset of $12,733$ single-pass transmembrane proteins newly uploaded in the UniProt website and with no restriction on the species. A greyed line highlights a lower performance of the maximal scoring segment using estimate $\hat{\sigma}$ compared with KD or TM scale.

6.3 Arbitrary prior distribution of the number of segments

We restricted the previous section (Section 6.2) to the particular framework of $N = 0$ or $N = 1$ segment as we focused on direct applications for the maximal score. However it follows a natural extension of the segment-based HMM (sb-HMM) to sequences containing multiple segments and/or multiple segment types. An interesting advantage of segment-based methods is their ability to control the prior distribution of the number of segments. The sb-HMM defined in the previous section is built under the assumption of a uniform prior distribution over the chosen segmentation space (so far $\mathcal{S}_n^{\{0\}}$, $\mathcal{S}_n^{\{1\}}$ or $\mathcal{S}_n^{\{0,1\}}$) leading to a very strong prior of the number of segments $N \in \{0, 1\}$ in favor of $N = 1$ when considering the segmentation space $\mathcal{S}_n^{\{0,1\}}$. We have seen in Section 6.2.3 and in particular in Equation 6.23 that one can choose a prior distribution $\mathbb{Q}$ for $N \in \{0, 1\}$. We now extend that idea to multiple segments and multiple segment types.

In this section functions ν and η_i are differently defined if compared to the previous section, however, we keep same notation for the sake of simplicity.

6.3.1 Method

Let us firstly detail the method in the particular framework of two segment types the we name respectively type-1 and type-2. The extension to multiple segment types will be developed at the end of the present section. We consider an HMM composed of a sequence of observed variables $X = (X_1, \ldots, X_n) \in \mathcal{X}^n$ and a hidden state sequence $S = (S_1, \ldots, S_n)$ where each S_i is a binary variable and takes its values in $\{1, 2\}$ such that $S_i = 1$ (respectively $S_i = 2$) if the i-th index is in a type-1 (respectively a type-2) segment. We denote by x_i the value taken by X_i and by $x = (x_1, \ldots, x_n)$ the vector of observed values taken by X. We assume that S is an homogeneous first order Markov chain i.e., for all $i \in \{2, \ldots, n\}$, $\mathbb{P}(S_i|S) = \mathbb{P}(S_i|S_{i-1})$ and we define the transition probability, for all $r, s \in \{1, 2\}$ to be $\mathbb{P}(S_i = s|S_{i-1} = r) = \pi(r, s)$. Furthermore we assume that the initial state probability is $\mathbb{P}(S_1 = s) = \pi(1, s)$ and we denote by $\nu(s, x_i)$ the emission probability at index i. Let $N \in \mathcal{N} = \{0, \ldots, \lceil \frac{n}{2} \rceil\}$ be the number of type-2 segments, we define

$$\mathbb{Q}(S, N, X = x) \stackrel{\text{def}}{=} \mathbb{P}(S, X = x|N)\mathbb{Q}(N) \tag{6.24}$$

where $\mathbb{Q}(N)$ is an arbitrary prior for N. Our main two goals are the computation of the posterior probability of N given by

$$\mathbb{Q}(N|X = x) = \frac{\mathbb{Q}(N, X = x)}{\mathbb{Q}(X = x)} \propto \mathbb{P}(X = x|N)\mathbb{Q}(N) \tag{6.25}$$

and individual posterior state probability

$$\begin{aligned}\mathbb{Q}(S_i|X=x) &= \sum_{k\geq 0}\mathbb{Q}(S_i, N=k|X=x)\\ &= \sum_{k\geq 0}\mathbb{Q}(S_i|X=x, N=k)\mathbb{Q}(N=k|X=x)\\ &= \sum_{k\geq 0}\mathbb{P}(S_i|X=x, N=k)\mathbb{Q}(N=k|X=x) \qquad (6.26)\end{aligned}$$

One can again take advantage of the FB algorithm for reducing the complexity to compute Equations (6.25) and (6.26).

Computation of $\mathbb{Q}(\mathbf{N}|\mathbf{X}=\mathbf{x})$. Recalling Equation (6.25), our goal is to compute, for all $k \in \mathcal{N}$,

$$\mathbb{P}(X=x|N=k) = \frac{\mathbb{P}(X=x, N=k)}{\mathbb{P}(N=k)} = \frac{\sum_S \mathbb{P}(S, X=x, N=k)}{\sum_S \mathbb{P}(S, N=k)}.$$

Let us introduce the function $\boldsymbol{\pi}$ defined on $\{1,2\}^2$ such that, for all $r, s \in \{1,2\}$, $\boldsymbol{\pi}(r,s) = \pi(r,s)z^{\mathbb{1}_{\{r=1,s=2\}}}$, note that we have

$$\sum_{k\geq 0}\mathbb{P}(X=x, N=k)z^k = \sum_S \prod_{i=1}^{n}\boldsymbol{\pi}(S_{i-1}, S_i)\eta_i(S_i) \qquad (6.27)$$

where $S_0 = 0$ by convention, z is a dummy variable and each η_i is defined over $\{1,2\}$ such that, for all $s \in \{1,2\}$, $\eta_i(s) = \nu(s, x_i)$ and

$$\sum_{k\geq 0}\mathbb{P}(N=k)z^k = \sum_S \prod_{i=1}^{n}\boldsymbol{\pi}(S_{i-1}, S_i). \qquad (6.28)$$

Therefore, one can apply a forward pass of the FB algorithm over the JT J represented in Figure 6.1b and the set of potentials respectively $\{\boldsymbol{\pi}\eta_i\}_{i=1,\dots,n}$ for Equation (6.27) and $\{\boldsymbol{\pi}\}_{i=1,\dots,n}$ for Equation (6.28). Let us make this point clearer by introducing the sequence $(N_1, \dots, N_n)$ such that $N_1 = \mathbb{1}_{\{S_1=2\}}$ and for $i = 2, \dots, n$, $N_i = N_{i-1} + \mathbb{1}_{\{S_{i-1}=1, S_i=2\}}$. Note that $N_n = N$. We define for all $i \in \{1, \dots, n\}$, the following polynomial forward messages

$$\boldsymbol{F}_i(S_i) \overset{\text{def}}{=} \sum_{k\geq 0}\mathbb{P}(X_1=x_1, \dots, X_i=x_i, S_i, N_i=k)z^k \qquad (6.29)$$

and

$$\boldsymbol{F}_i^0(S_i) \overset{\text{def}}{=} \sum_{k\geq 0}\mathbb{P}(S_i, N_i=k)z^k \qquad (6.30)$$

which can be recursively computed with a forward pass in $\mathcal{O}(n \times |\mathcal{N}|)$ in time and, at the end of the recursion, we get

$$\sum_{s\in\{1,2\}}\boldsymbol{F}_n(s) = \sum_{k\geq 0}\mathbb{P}(X=x, N=k)z^k \quad \text{and} \quad \sum_{s\in\{1,2\}}\boldsymbol{F}_n^0(s) = \sum_{k\geq 0}\mathbb{P}(N=k)z^k.$$

In order to further reduce the time complexity, one can appropriately choose a maximal number of segments $k_{\max}$ and replace the conventional product of polynomials by the multiplicative law denoted $\star$ defined as the conventional product with an additional truncation at degree $k_{\max}$ for a recursive implementation in $\mathcal{O}(n \times k_{\max})$ in time to compute

$$\sum_{s\in\{1,2\}} \boldsymbol{F}_n(s) = \sum_{k=0}^{k_{\max}} \mathbb{P}(X = x, N = k) z^k \quad \text{and} \quad \sum_{s\in\{1,2\}} \boldsymbol{F}_n^0(s) = \sum_{k=0}^{k_{\max}} \mathbb{P}(N = k) z^k.$$

Finally, for each $k \in \{0, \ldots, k_{\max}\}$, $\mathbb{P}(X = x|N = k)$ is given by

$$\frac{\mathbb{P}(X = x, N = k)}{\mathbb{P}(N = k)} = \frac{\left[\sum_{s\in\{1,2\}} \boldsymbol{F}_n(s)\right]_{z^k}}{\left[\sum_{r\in\{1,2\}} \boldsymbol{F}_n^0(r)\right]_{z^k}}$$

where $\left[\,\cdot\,\right]_{z^k}$ denotes the extraction of the k-th coefficient in a polynomial.

Computation of $\mathbb{Q}(\mathbf{S}|\mathbf{X} = \mathbf{x})$. Returning to Equation (6.26), our goal is to infer $\mathbb{P}(S_i|X = x, N = k)$ for all $k \leq k_{\max}$. We define the following polynomial backward messages, for all $i \in \{1, \ldots, n\}$,

$$\boldsymbol{B}_i(S_i) \stackrel{\text{def}}{=} \sum_{k\geq 0} \mathbb{P}(X_{i+1} = x_{i+1}, \ldots, X_n = x_n, N - N_i = k|S_i) z^k$$

Similarly, polynomial backward can be write as

$$\boldsymbol{B}_i(S_i) = \sum_{S_{i+1}} \cdots \sum_{S_n} \prod_{j=i+1}^{n} \boldsymbol{\pi}(S_{j-1}, S_j) \eta_j(S_j)$$

and therefore, one can compute them in $\mathcal{O}(n \times k_{\max})$ in time using a backward pass of the FB algorithm over J with potentials $\{\boldsymbol{\pi}\eta_i\}_{i=1,\ldots,n}$ and replacing the conventional product by $\star$ defined as the conventional product with an additional truncation step at degree $k_{\max}$.

Finally, note that, by definition, we have

$$\mathbb{P}(S_i = s|X = x, N = k) = \frac{\left[\boldsymbol{F}_i(s) \star \boldsymbol{B}_i(s)\right]_{z^k}}{\sum_{r\in\{1,2\}} \left[\boldsymbol{F}_i(r) \star \boldsymbol{B}_i(r)\right]_{z^k}}$$

which is computed, in addition to the forward recursion for polynomial forward messages, using a single backward recursion for all $i \in \{1, \ldots, n\}$ and all $k \in \mathcal{N}$ leading to two forward and one backward recursions for Equation (6.26), hence a resulting complexity for $\mathbb{Q}(S|X = x)$ of the order of $\mathcal{O}(n \times k_{\max})$ in time.

Remark about real-valued polynomial potentials versus real-valued potentials. Some alternative FB recursions over real-valued potentials are applicable for computing the same quantities. Let us mention one of them of lowest time complexity before pointing at advantages of a recursion over polynomial potentials.

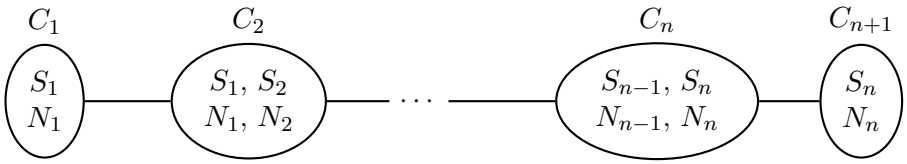


Figure 6.9: Junction-tree associated with a recursion over real-valued potentials.

We introduce a set of n variables $N = \{N_1, \ldots, N_n\} \in \{0, \ldots, k_{\max}\}^n$ such that, for all $i \in \{1, \ldots, n\}$, N_i denotes the number of type-2 segments up to sequence index i. Let us define, for all $i \in \{1, \ldots, n\}$, the following forward and backward messages

$$F_i^{\text{real}}(S_i, N_i) = \mathbb{P}(X_1 = x_1, \ldots, X_i = x_i, S_i, N_i)$$

and

$$B_i^{\text{real}}(S_i, N_i) = \mathbb{P}(X_{i+1} = x_{i+1}, \ldots, X_n = x_n | S_i, N_i),$$

which can be recursively implemented with the FB algorithm over the JT pictured in Figure 6.9 and real-valued potentials. Entering the evidence $\{N_n = k\}$ (or equivalently applying the constraint, for $s \in \{1, 2\}$, for all $\ell \neq k$, $B_n^{\text{real}}(s, \ell) = 0$ and $B_n^{\text{real}}(s, k) = 1$ during the initialization of the backward pass), one can compute $\mathbb{P}(S_i = s, X = x, N_n = k)$ in $\mathcal{O}(n \times k_{\max})$ time complexity for a chosen $k \in \{0, \ldots, k_{\max}\}$. However, one must repeat a backward recursion with the right constraint, for each $k' \in \{0, \ldots, k_{\max}\}$, for all $\ell \neq k'$, $B_n^{\text{real}}(s, \ell) = 0$ and $B_n^{\text{real}}(s, k') = 1$ in order to compute $\mathbb{Q}(S|X = x)$ leading to a total complexity of the order of $\mathcal{O}(n \times {k_{\max}}^2)$ in time. Therefore a recursion over polynomial potentials allows not only for a more compact implementation but also for a complexity reduction as well as a natural extension to multiple segment types as detailed in the next paragraph.

Extension to multivariate polynomials. One can extend the aforementioned method to multiple segment types with multivariate polynomials. Let us consider D segment types, each S_i takes its values in $\{1, \ldots, D\}$ and the transition probability π is defined on $\{1, \ldots, D\}^2$ such that, for all $r, s \in \{1, \ldots, D\}$, $\pi(r, s) = \mathbb{P}(S_i = s | S_{i-1} = r)$. Let $N^1, \ldots, N^D$ be D variables such that, for all $j \in \{1, \ldots, D\}$, N^j denotes the number of type-j segment(s), we introduce D sets $\{N_i^1\}_{i=1,\ldots,n}$, …, $\{N_i^D\}_{i=1\ldots,n}$, such that, for $j \in \{1, \ldots, D\}$, $N_1^j = \mathbb{1}_{\{S_1 = j\}}$ and for $i \in \{2, \ldots, n\}$, $N_i^j = N_{i-1}^j + \mathbb{1}_{\{S_{i-1} \neq j, S_i = j\}}$. Note that $N^1 = N_n^1, \ldots, N^D = N_n^D$. Our goal is to compute

$$\mathbb{Q}(N^1, \ldots, N^D | X = x) = \frac{\mathbb{P}(X = x, N^1, \ldots, N^D)}{\mathbb{P}(N^1, \ldots, N^D)} \mathbb{Q}(N^1, \ldots, N^D)$$

where $\mathbb{Q}(N^1, \ldots, N^D)$ is an arbitrary prior for $N = \{N^1, \ldots, N^D\}$ and

$$\mathbb{Q}(S_i | X = x) = \sum_{k_1 \geq 0} \cdots \sum_{k_D \geq 0} \mathbb{P}(S_i | X = x, N^1 = k_1, \ldots, N^D = k_D) \times \mathbb{Q}(N^1 = k_1, \ldots, N^D = k_D | X = x).$$

Let us define the following polynomial forward and backward messages:

$$\boldsymbol{F}_i(S_i) \stackrel{\text{def}}{=} \sum_{k_1\geq 0} \dots \sum_{k_D\geq 0} \mathbb{P}(X_1 = x_1, \dots, X_i = x_i, S_i, N_i^1 = k_1, \dots, N_i^D = k_D) z_1^{\ell_1} \dots z_D^{k_D},$$

and

$$\boldsymbol{F}_i^0(S_i) \stackrel{\text{def}}{=} \sum_{k_1\geq 0} \dots \sum_{k_D\geq 0} \mathbb{P}(S_i, N_i^1 = k_1, \dots, N_i^D = k_D) z_1^{k_1} \dots z_D^{k_D}$$

where $z_1, \dots, z_D$ are D dummy variables, note that we have

$$\frac{\mathbb{P}(X = x, N^1 = k_1, \dots, N^D = k_D)}{\mathbb{P}(N^1 = k_1, \dots, N^D = k_D)} = \frac{\left[\sum_{s\in\{1,\dots,D\}} \boldsymbol{F}_n(s)\right]_{z_1^{k_1}\dots z_D^{k_D}}}{\left[\sum_{r\in\{1,\dots,D\}} \boldsymbol{F}_n^0(r)\right]_{z_1^{k_1}\dots z_D^{k_D}}}. \tag{6.31}$$

Furthermore, defining polynomial backward messages to be

$$\boldsymbol{B}_i(S_i) \stackrel{\text{def}}{=} \sum_{k_1\geq 0} \dots \sum_{k_D\geq 0} \mathbb{P}(X_{i+1} = x_{i+1}, \dots, X_n = x_n, N^1 - N_i^1 = k_1, \dots, N^D - N_i^D = k_D | S_i) \times z_1^{k_1} \dots, z_D^{k_D},$$

note that we have

$$\mathbb{P}(S_i = s | X = x, N^1 = k_1, \dots, N^D = k_D) = \frac{\left[\boldsymbol{F}_i(s) \star \boldsymbol{B}_i(s)\right]_{z_1^{k_1}\dots z_D^{k_D}}}{\sum_{r\in\{1,\dots,D\}} \left[\boldsymbol{F}_i(r) \star \boldsymbol{B}_i(r)\right]_{z_1^{k_1}\dots z_D^{k_D}}} \tag{6.32}$$

Similarly with the previous section, introducing the transition function $\boldsymbol{\kappa}$ defined on $\{1,2\}^2$ such that for all $r, s \in \{1,2\}$, $\boldsymbol{\kappa}(r,s) = \pi(r,s)\prod_{j=1}^D z_j^{\mathbb{1}_{\{r\neq j, s=j\}}}$, forward and backward messages of the form $\boldsymbol{F}_i$, $\boldsymbol{F}_i^0$ and $\boldsymbol{B}_i$ can be recursively computed over the JT represented in Figure 6.1b respectively with a forward pass of the FB algorithm and potentials $\{\boldsymbol{\kappa}\eta_i\}_{i=1,\dots,n}$, a forward pass and potentials $\{\boldsymbol{\kappa}\}_{i=1,\dots,n}$ and a backward pass and potentials $\{\boldsymbol{\kappa}\eta_i\}_{i=1,\dots,n}$. Replacing the conventional product by $\star$, the conventional product with truncation at degree $k_{\max}$, each pass is of the order of $\mathcal{O}(n \times {k_{\max}}^D)$ for computing Equation (6.31) and (6.32) with $k_1 + \dots + k_D \leq k_{\max}$.

6.3.2 Applications

This application section is restricted to $D = 2$ segment types. This brief section is still restricted to a limited selection of simulated datasets and needs to be densified with extensive simulation schemes and real datasets as well as comparisons with other methods. These extensions are in progress and will appear in future works.

Let $X = \{X_1, \dots, X_n\} \in \mathcal{X}^n$ be the set of observed variable, we denote by x_i the observed value taken by X_i. Let $S = (S_1, \dots, S_n) \in \mathcal{S}_n^{\{0,\dots,k_{\max}\}}$ be the state sequence space where $\mathcal{S}_n^{\{0,\dots,k_{\max}\}}$ is the set of sequences of length n containing at most $k_{\max}$ segment(s) where $k_{\max}$ is arbitrarily chosen according to the context. We

consider solely two types of segments, hence, each S_i takes its value in $\{1,2\}$ and $S_i = 1$ (respectively $S_i = 2$) if the i-th index is in a type-1 (respectively in a type-2) segment. Emission probabilities are defined on $\{1,2\}\times\mathcal{X}$ such that, for all $s \in \{1,2\}$, for all $a \in \mathcal{X}$, $\mu(s,a) = q_{s-1}(a)$ where $q_0(a)$ (respectively $q_1(a)$) is the proportion of letter a in type-1 (respectively type-2) segments. In order to make the notation simpler, we assume that $\mathcal{X} = \{1,\dots,d\}$ and we denote the vectors $(q_0(a))_{a=1,\dots,d}$ and $(q_1(a))_{a=1,\dots,d}$ respectively simply by q_0 and q_1.

We assume that $\mathbb{P}(S)$ is uniform over $\mathcal{S}_n^{\{0,\dots,k_{\max}\}}$ and therefore we replace functions π and $\boldsymbol{\pi}$ in previous sections by the transition functions τ and $\boldsymbol{\tau}$ such that, for all $r,s \in \{1,2\}$, $\tau(r,s) = 1$ and $\boldsymbol{\tau}(r,s) = z^{\mathbb{1}_{\{s-r=1\}}}$. Note that τ simply ensures a uniform distribution over the state sequence space and is solely defined up to a coefficient of proportionality with no consequence on the result as we solely need to compute ratios $\big[\mathbb{P}(X=x,N)/\mathbb{P}(N)\big]$ and $\big[\mathbb{P}(S_i,X=x,N)/\mathbb{P}(X=x,N)\big]$.

In particular, returning to the forward messages defined in Equation (6.30) and recursively implemented such that, for $s \in \{1,2\}$,

$$\boldsymbol{F}_1^0(s) = \boldsymbol{\pi}(1,s) \quad \text{and for } i \in \{2,\dots,n\} \quad \boldsymbol{F}_i^0(s) = \sum_{r\in\{1,2\}} \boldsymbol{F}_{i-1}^0(r)\boldsymbol{\pi}(r,s). \tag{6.33}$$

Now, defining recursively forward message, for $i = 1,\dots,n$, $\boldsymbol{F}_i^{0,\tau}$ as in Equation (6.33) where $\boldsymbol{\pi}$ is replaced by $\boldsymbol{\tau}$, we get

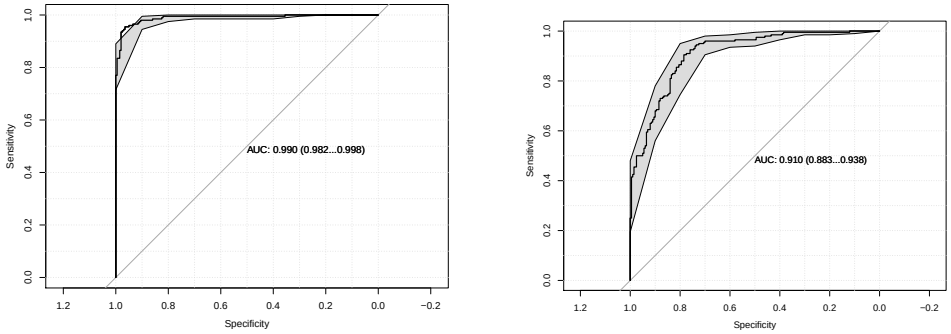
$$\sum_{s\in\{1,2\}} \boldsymbol{F}_n^{0,\tau}(s) = \sum_{k=0}^{k_{\max}} |\mathcal{S}_n^{\{k\}}| z^k = \sum_{k=0}^{k_{\max}} \binom{n+1}{2k} z^k$$

where $S_n^{\{k\}}$ is the set of sequences containing exactly k segments. Consequently we have

$$\Big[\sum_{s\in\{1,2\}} \boldsymbol{F}_n^{0,\tau}(s)\Big]_{z^k} = |\mathcal{S}_n^{\{k\}}| = \binom{n+1}{2k} \propto \mathbb{P}(N=k). \tag{6.34}$$

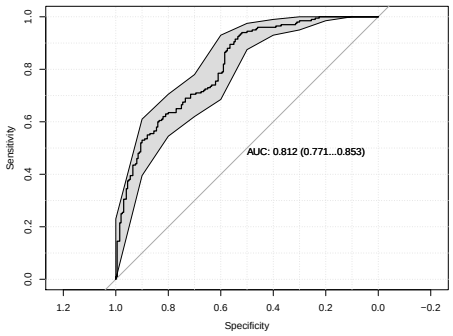
For this whole application section, we assume that $\mathbb{Q}$ is a uniform prior for $N \in \{0,\dots,k_{\max}\}$.

Posterior probability of N in various simulation schemes. We simulated four sets of 200 sequences over the alphabet $\mathcal{X} = \{0,1\}$. In the first (respectively second, third and fourth) set, sequences are simulated with $N^* = 0$ (respectively $N^* = 1$, $N^* = 2$ and $N^* = 3$) type-2 segment(s) of length $L_{\text{seg}} = 30$ letters each. In the first set, each sequence is of length 200 letters. In order to keep a constant length ratio between type-1 and type-2 segments for future interpretation, sequence lengths are respectively 200, 400 and 600 letters in the second, third and fourth set. Segment localizations are randomly simulated with a minimum of 30 letters between each segment. We repeated simulations for $L_{\text{seg}} = 60$ and $L_{\text{seg}} = 90$ letters. For all simulation schemes q_0 is uniform over $\mathcal{X}$ and q_1 varies.

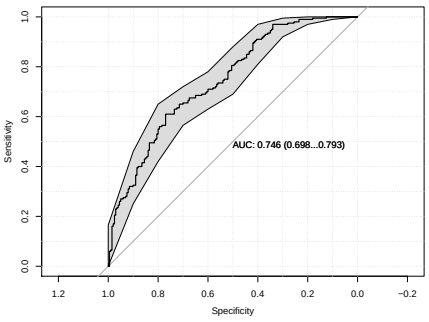


(a) BF_0

(b) BF_1



(c) BF_2



(d) BF_3

Figure 6.10: ROC curves and associated AUC for Bayes factors BF_k for $k \in \{0, \ldots, 3\}$, atypical segments length of 30 letters, $q_1 = (0.15, 0.85)$ and $k_{\max} = 15$.

Let us propose hypothesis tests in order to question whether sequences contain k type-2 segments ($\mathbb{H}_0^k$) or not ($\mathbb{H}_1^k = \cup_{j \neq k} \mathbb{H}_0^k$) based on the Bayes factor:

$$\mathrm{BF}_k = \frac{\mathbb{Q}_{\mathbb{H}_0^k}(N = k | X = x)}{\mathbb{Q}_{\mathbb{H}_1^k}(N = k | X = x)}.$$

Results are displayed in Figure 6.10 and Tables 6.10 and 6.11. Figure 6.10 is a graphical representation of ROC curves computed for Bayes factors BF_k for $k = 0, \ldots, 3$ using $L_{\mathrm{seg}} = 30$, $q_1 = (0.15, 0.85)$ and $k_{\max} = 15$. Tables 6.10 (respectively Tables 6.11) gives AUC for Bayes factors BF_k for $k = 0, \ldots, 3$ computed with various q_1, $k_{\max} = 5$ or $k_{\max} = 15$ and $L_{\mathrm{seg}} = 30$ (respectively $k_{\max} = 15$ and $L_{\mathrm{seg}} = 30$, $L_{\mathrm{seg}} = 60$ or $L_{\mathrm{seg}} = 90$ letters). As expected we can see that AUCs augment with an increasing difference between $q_1(0)$ and $q_1(1)$ as well as a ratio between type-2 and type-1 segments lengths tending towards one. Furthermore AUCs are better when setting $k_{\max}$ closer to the true number of segments. These results suggest a right behavior of the estimator, however we still need to perform further theoretical and asymptotical analyses.

In the second simulation scheme, we simulated four sets of 100 sequences of length $n = 150$ (respectively $n = 300$, $n = 600$ and $n = 3000$) letters. Each

q_1	$\begin{bmatrix}0.01\\0.99\end{bmatrix}$	$\begin{bmatrix}0.05\\0.95\end{bmatrix}$	$\begin{bmatrix}0.15\\0.85\end{bmatrix}$	$\begin{bmatrix}0.20\\0.80\end{bmatrix}$	$\begin{bmatrix}0.30\\0.70\end{bmatrix}$
BF_0	>0.999	>0.999	0.989	0.958	0.789
BF_1	0.994	0.984	0.910	0.799	0.564
BF_2	0.967	0.946	0.812	0.716	0.581
BF_3	0.916	0.862	0.746	0.685	0.525

(a) $k_{\max} = 15$

q_1	$\begin{bmatrix}0.01\\0.99\end{bmatrix}$	$\begin{bmatrix}0.05\\0.95\end{bmatrix}$	$\begin{bmatrix}0.15\\0.85\end{bmatrix}$	$\begin{bmatrix}0.20\\0.80\end{bmatrix}$	$\begin{bmatrix}0.30\\0.70\end{bmatrix}$
BF_0	>0.999	>0.999	0.991	0.963	0.808
BF_1	0.994	0.988	0.939	0.849	0.653
BF_2	0.970	0.953	0.849	0.760	0.630
BF_3	0.943	0.909	0.821	0.771	0.607

(b) $k_{\max} = 5$

Table 6.10: AUC computed for BF_k, $k = 0, \ldots, 3$, with $L_{\text{seg}} = 30$ letters, various q_1 and $k_{\max} = 15$ (Table 6.10a) or $k_{\max} = 5$ (Table 6.10b).

q_1	$\begin{bmatrix}0.15\\0.85\end{bmatrix}$	$\begin{bmatrix}0.20\\0.80\end{bmatrix}$	$\begin{bmatrix}0.30\\0.70\end{bmatrix}$
BF_0	0.989	0.958	0.789
BF_1	0.910	0.799	0.564
BF_2	0.812	0.716	0.581
BF_3	0.746	0.685	0.525

(a) $L_{\text{seg}} = 30$ letters

q_1	$\begin{bmatrix}0.15\\0.85\end{bmatrix}$	$\begin{bmatrix}0.20\\0.80\end{bmatrix}$	$\begin{bmatrix}0.30\\0.70\end{bmatrix}$
BF_0	>0.999	0.997	0.971
BF_1	0.975	0.955	0.861
BF_2	0.921	0.887	0.799
BF_3	0.844	0.791	0.716

(b) $L_{\text{seg}} = 60$ letters

q_1	$\begin{bmatrix}0.15\\0.85\end{bmatrix}$	$\begin{bmatrix}0.20\\0.80\end{bmatrix}$	$\begin{bmatrix}0.30\\0.70\end{bmatrix}$
BF_0	>0.999	>0.999	0.996
BF_1	0.970	0.963	0.927
BF_2	0.926	0.893	0.824
BF_3	0.851	0.810	0.711

(c) $L_{\text{seg}} = 90$ letters

Table 6.11: AUC computed for BF_k, $k = 0, \ldots, 3$ with various q_1, $k_{\max} = 15$ and $L_{\text{seg}} = 30$ (Table 6.11a), $L_{\text{seg}} = 60$ (Table 6.11b) or $L_{\text{seg}} = 90$ (Table 6.11c).

sequence contains one type-2 segment simulated with $q_1 = (0.15, 0.85)$. Each type-2 segment is of length equal a third of that of the sequence and positioned at the middle of the sequence. Figure 6.11 represents the mean of $\mathbb{Q}(N|X = x, k_{\max} = 15)$. As expected, at constant length ratio, as the length of the sequence grows, the distribution $\mathbb{Q}(N|X = x, k_{\max} = 15)$ tends toward a Dirac distribution at the true number of segment $N^* = 1$. These results suggest again a right behavior of the estimator and we need to pursue analyses in various simulation schemes.

Remark: Note that, as $\mathbb{P}$ is uniform over $\mathcal{S}_n^{\{0,\ldots,k_{\max}\}}$, quantities $\mathbb{P}(N = k_{\max}|X = x) >> \mathbb{P}(N = k|X = x)$ for all $k < k_{\max}$ in all simulation schemes (results not shown) due to the strong prior pointed at in Equation 6.34.

Posterior probability of S. In the last simulation scheme, we simulated two sequences with $q_1 = (0.15, 0.85)$. The first (respectively second) sequence is of length 600 (respectively 630) letters and contain two (respectively three) type-2 segments. For both sequences we compute, for all $i \in \{1, \ldots, n\}$

$$\mathbb{Q}(S_i = 2|X = x, k_{\max} = 15) = \sum_{k=0}^{k_{\max}} \mathbb{P}(S|X = x, N = k)\mathbb{Q}(N = k|X = x)$$

and

$$\mathbb{P}(S_i = 2|X = x, k_{\max} = 15) = \sum_{k=0}^{k_{\max}} \mathbb{P}(S|X = x, N = k)\mathbb{P}(N = k|X = x).$$

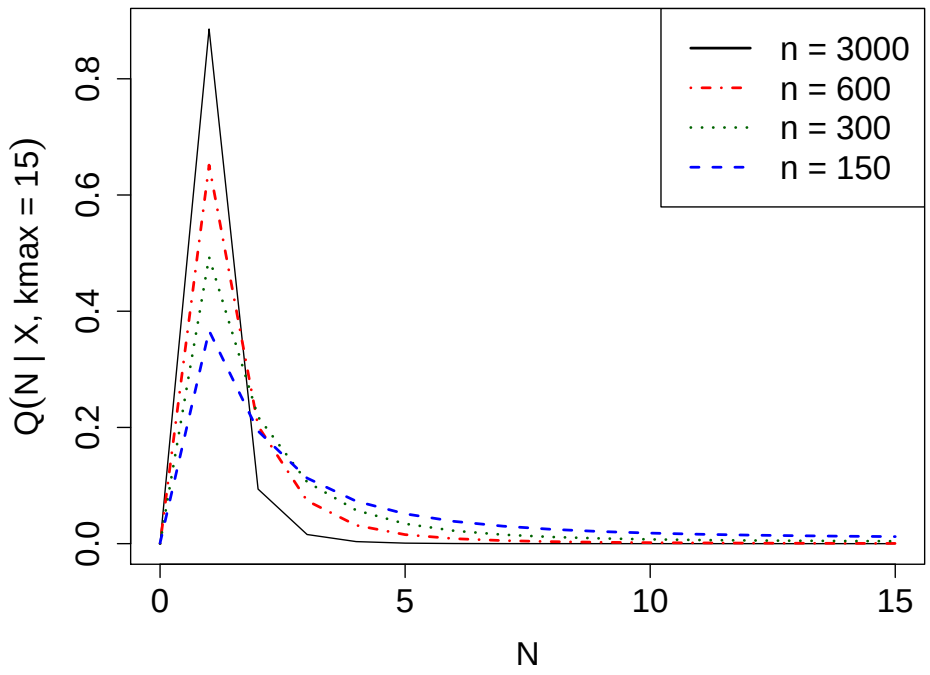
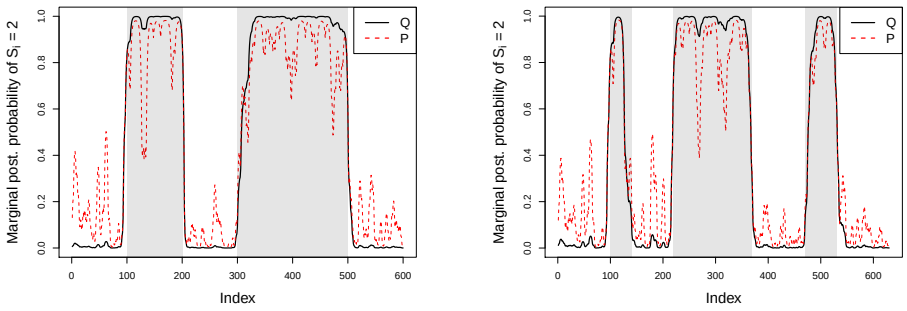


Figure 6.11: Mean of $\mathbb{Q}(N|X = x, k_{\max} = 15)$ for sets of sequences of length $n = 150$ to $n = 3000$ letters containing each one atypical segment of length $n/3$.



(a) First sequence (b) Second sequence

Figure 6.12: Marginal posterior probabilities $\mathbb{Q}(S_i = 2|X, k_{\max} = 15)$ (solid black line) and $\mathbb{P}(S_i = 2|X, k_{\max} = 15)$ (dashed red line) for the two simulated sequences containing respectively two and three type two segments. The true positions of type two segments are delimited by the greyed regions.

Results are shown in Figure 6.12 where $\mathbb{Q}(S_i = 2|X = x, k_{\max} = 15)$ (respectively $\mathbb{P}(S_i = 2|X = x, k_{\max} = 15)$) is drawn in solid black line (respectively dashed red line) and the true position of type-2 segments are delimited by greyed regions. These results suggest a better performance to highlight type-2 segments with $\mathbb{Q}$. Indeed, as $\mathbb{P}$ is uniform over $\mathcal{S}_n^{\{0,\dots,k_{\max}\}}$, such that $\mathbb{P}(N = k) \propto |\mathcal{S}_n^{\{k\}}|$, this strong prior tends to favor sequences with an increasing number of segments. Let us introduce the distribution $\mathcal{P}(S_i = s) = \mathbb{1}_{S_i=s}$ for $s \in \{1, 2\}$, the Kullback-Leibler divergence from a distribution $\mathcal{Q}(S|X = x)$ to $\mathcal{P}(S)$ is defined as

$$\mathrm{KL}(\mathcal{P}||\mathcal{Q}) = \sum_{i=1}^{n} \mathcal{P}(S_i) \log\left(\frac{\mathcal{P}(S_i)}{\mathcal{Q}(S_i|X = x)}\right) = \sum_{i=1}^{n} \sum_{s\in\{1,2\}} -\mathbb{1}_{\{S_i=s\}} \log \mathcal{Q}(S_i = s|X = x).$$

The Kullback-Leibler divergence from $\mathcal{P}$ to $\mathbb{P}$ and from $\mathcal{P}$ to $\mathbb{Q}$ compute from our datasets is: $\mathrm{KL}(\mathcal{P}||\mathbb{Q}) = 34.47$ (respectively $\mathrm{KL}(\mathcal{P}||\mathbb{Q}) = 62.09$) and $\mathrm{KL}(\mathcal{P}||\mathbb{P}) = 99.26$ (respectively $\mathrm{KL}(\mathcal{P}||\mathbb{P}) = 114.59$) for the first (respectively the second) sequence. Thus the relative entropy of $\mathbb{Q}$ is lower than that of $\mathbb{P}$ with respect to $\mathcal{P}$.

Let us finally recall the definition of the Brier score of a distribution $\mathcal{Q}(S|X = x)$ to be

$$\mathrm{BS}_{\mathcal{Q}} = \frac{1}{n} \sum_{i=1}^{n} \left(\mathcal{Q}(S_i = 2|X = x) - \mathbb{1}_{S_i=2}\right)^2,$$

we have $\mathrm{BS}_{\mathbb{Q}} = 0.017$ (respectively $\mathrm{BS}_{\mathbb{Q}} = 0.031$) and $\mathrm{BS}_{\mathbb{P}} = 0.040$ (respectively $\mathrm{BS}_{\mathbb{P}} = 0.048$) for the first (respectively the second) sequence suggesting better predictive performances of $\mathbb{Q}$ versus $\mathbb{P}$ for marginal posterior state probabilities.

Remark. This whole application section constitutes a first step of evaluation suggesting a right behavior of estimators $\mathbb{Q}(N|X = x)$ and $\mathbb{Q}(S|X = x)$. However we

still need to densify that section with both theoretical results and extensive simulation schemes as well as applications over real datasets. We also need to compare our results with those of conventional HMMs, in particular, evaluate the statistical interest (or not) of controlling the prior number of segments versus prior transition probabilities in level-based methods.

6.4 Conclusion

In this chapter, we proposed two contributions motivated by the dual relation between the maximal score and a constrained HMM as proved by Mercier and Nuel (2021) when using SF $\sigma = \log(q_1/q_0)$, where q_1 (respectively q_0) is the proportion of letters in atypical (respectively non-atypical) segments.

Firstly we proposed an unsupervised method for the statistical learning of a SF for the maximal score based on a constrained segment-based HMM. The HMM framework allows one to apply the FB algorithm to compute estimates of the SF with the EM algorithm in linear time complexity in the length of the longest sequence as well as estimates of CIs for the SF in quadratic time in the parameter dimension using, for example, the method proposed in Lefebvre and Nuel (2018). Furthermore, for a fixed parameter q (for instance the estimate $\hat{q}$), one can compute posterior probabilities of interest such as posterior sequence state probabilities, posterior probabilities of start and stop positions of the atypical segment, posterior probability that a sequence contain an atypical segment as well as the most probable state sequence in linear time complexity in the length of the sequence. We illustrated the interest of our method with simulated datasets and a real dataset of single-pass TM proteins downloaded from UniProt website.

Estimates on simulated dataset with random sequence lengths between 150 and 1000 letters and random segment length at random positions were unbiased both for θ and σ. Furthermore coverage probabilities estimated from datasets of 200 sequences with true parameter $\theta^* = (0.00, 0.00, 0.00, 1.37, -0.56, 0.36)$ were the expected ones in case of a length ratio of $40/150 \approx 0.27$ but where slightly biased for a length ratio of $40/1000 = 0.04$ suggesting a sensitivity of the method to length ratios between atypical and non-atypical segments. Using estimate $\hat{q}$ to compute the posterior probability that a sequence contains an atypical segment (H_0) or not (H_1) over a dataset of sequences composed of 30% of sequences with no atypical segment led to an AUC of 0.9638 $[0.9495; 0.9781]$ using the Bayes factor $\mathrm{BF} = \mathbb{Q}_{H_0}(N = 1|X = x, \hat{q})/\mathbb{Q}_{H_1}(N = 1|X = x, \hat{q})$ where $\mathbb{Q}(N)$ is a uniform prior over $\{0, 1\}$. Furthermore the sensitivity and specificity of the maximal scoring segment computed with $\hat{\sigma} = \log(\hat{q}_1/\hat{q}_0)$ were 0.9551 and 0.9787 respectively. However, we need to pursue analyses on simulated datasets with a variety of simulation schemes, especially with various sequence lengths and length ratios, various parameters θ and various dimensions for θ (i.e. cardinal of the alphabet).

Analyses on a real dataset of TM proteins showed other strengths and weaknesses of the method in a context of a high-dimensional parameter (18 dimensions for amino-acid sequences). First of all, as we use the EM algorithm for computing the MLE, the method is highly sensitive to its initialization, hence it may lead to biased estimates in difficult frameworks such as low length ratios between atypical and non-atypical segments. Hence, estimates were computed on a dataset restricted to sequences of length at most 500 amino acids. Secondly some biased estimates for θ^1 showed another weakness in case of an extreme rarity of some amino acids in atypical segments. A solution for overcoming that issue could be an arbitrary choice of a fixed (extremely low) frequency for these amino-acids in order to solely compute

estimates for more frequent ones. We obtained a very high specificity but a lower sensitivity of the maximal scoring segment computed with $\hat{\sigma}$ in comparison with the Kyte and Doolittle scale and Zhao's TM tendency scale. This can be a consequence of a unique temperature T rendering both models defined Equation (6.13) and (6.14) equivalent using SF $\sigma = f/T$ with $f = T \log(q_1/q_0)$. Empirical SFs are indeed more flexible and can be adapted in order to better highlight a particular feature of interest. However, we think that the unsupervised nature of our method could be of great interest in a domain in which empirical SFs are difficult to establish and/or the feature of interest is insufficiently known. An article is in preparation for which we need to develop theoretical results, densify the application section with a variety of simulation schemes and apply the method to real datasets in a domain in need of unsupervised methods.

Secondly we extended the constrained segment-based HMM to multiple segments and multiple segment types and in particular, we proposed a method for allowing for an arbitrary prior distribution $\mathbb{Q}$ of the number of segments. We showed the interest of the model through the computation of $\mathbb{Q}(N|X = x)$ and $\mathbb{Q}(S|X = x)$ in $\mathcal{O}(n \times k_{\max})$ time complexity when using polynomial potentials where $k_{\max}$ is an arbitrary prior on the maximal number of segments in the sequence. We compare our results with $\mathbb{P}(S|X = x)$ where $\mathbb{P}$ is uniform over the state sequence space. Preliminary results on simulation schemes show a right behavior of the estimator $\mathbb{Q}(N|X = x)$ whereas $\mathbb{P}$ tends to overestimate $\mathbb{P}(N = k_{\max}|X = x)$. Furthermore first results suggest a better performance for segments localization when using $\mathbb{Q}$ rather than $\mathbb{P}$. A first perspective of this work is to density the application section in a variety of simulation schemes as well applying our model to real datasets. Furthermore, an important other perspective is a comparison with classical level-based HMMs. Preliminary results are quite disappointing with sequences containing segments of high heterogeneity but seems to be promising in case of lower heterogeneity. Note that we are currently limited to sequences of reasonable length as we are still investigating a computational trick to avoid underflow issues with our model. Indeed, as we deal with polynomials with coefficient of extremely various ranges, no trivial method exists for dealing with computational underflow and we are still working on it. Therefore the aforementioned perspectives remain very limited and extensive comparison will be done once the underflow issues are overcome.

Part III

LynchRisk: a pedigree-based model for the Lynch syndrome

Chapter 7

Epidemiological and clinical context

Sommaire

The third part of the thesis is devoted to the development of a new pedigree-based model for computing probabilities of genetic predisposition and cancer risks in the framework of the Lynch syndrome (LS). Most consultations in genetic counseling are related to the breast/ovarian syndrome with 54 936 consultations in 2017 in France and to the Lynch syndrome with 8 020 consultations in 2017 in France (`www.e-cancer.fr`). In the first and present chapter, we start with a review about the Lynch syndrome in order to introduce important notions to understand the construction of the model developed in Chapter 8. This chapter is organized as follows: in Section 7.1 we start with an introduction to cancer and cancer genetics in order to highlight the notion of genetic predisposition to cancer. The Lynch syndrome is introduced in Section 7.2 with its definition and main epidemiological data. In Section 7.3 we detail current tools for guiding clinicians and genetic counselors in their risk assessment and decision making. We start with first guidelines based on qualitative data, we pursue with clinical and biological data associated with the Lynch

syndrome before detailing main current mathematical models along with their advantages and their limitations. We finally expose our motivations for proposing a new pedigree-based model in this context. The present chapter is not intended for future publication but it is essential for acquiring an overview of the framework the model developed in the next chapter is built in.

7.1 Cancer and carcinogenesis

7.1.1 Cancer epidemiology

The word *cancer* comes from *"karknios"* (crabe in ancient Greek), a name first given by Hippocrate for the visual appearance of some tumors with a round central corp and spreading veins. A cancer is a disease characterized by an uncontrolled cell growth and proliferation. There exists many types of cancers and a wide heterogeneity regarding their localization, biological and histological profile, risk factors, prognostic, etc. and we often speak about cancers in plural. Cancer is a multifactorial genetic disease in the sense that it originally comes from an accumulation of mutations in genes involved in cell growth and division and multiple causes are involved as well as Gene × Gene and Gene × Environment interactions.

Cancer is the worldwide second leading cause of death and the first in men and the second in women in France. According to WHO[1] (the World Health Organization) it is responsible for an estimated 9.6 million worldwide deaths in 2018. The worldwide most common cancers in 2020 are breast (2.26 million cases, 11.7%), lung (2.21 million cases, 11.4%), colon-rectum (1.93 million cases, 10.0%), prostate (1.41 million cases, 7.3%) and stomach (1.09 million cases, 5.06%) and the leading cause of death by cancer are lung (1.80 million, 18.2%), colon-rectum (0.93 million, 9.5%), liver (0.83 million, 8.4%) and stomach (0.77 million, 7.8%) according to IARC[2] (the International Agency for Research on Cancer).

Santé Publique France[3] and INCa[4] (*Institut National du Cancer*), in partnership with the biostatistics and bioinformatics department of civil hospices in Lyon and Francim (*Réseau français des registres des cancers*), regularly publish an overview about cancer research in France as well as main epidemiological indicators. Defossez et al. (2019) estimated 382 000 new cases (204 600 for men and 177 400 for women) and 157,400 death by cancer (89 600 for men and 67 800 for women) in France in 2018. Incidence and mortality rates are usually age-standardized in order to take into account the increasing population life expectancy and keep comparable data. Estimates of new cases and death as well as raw and age-standardized annual incidence and mortality rates for 2015 (prostate) or 2018 (other localizations) are given in Table 7.1 for the main three cancers in France per sex (Defossez et al., 2019).

[1] https://www.who.int
[2] https://gco.iarc.fr/today/home
[3] https://www.santepubliquefrance.fr
[4] https://www.e-cancer.fr/

Localisation	Number of new cases	Number of death	AIR / 100 000 p-y Raw	AIR / 100 000 p-y Standardized	AMR / 100 000 p-y Raw	AMR / 100 000 p-y Standardized
All localizations	204 583	89 621	649.5	330.2	284.5	123.8
Prostate	50 430	8 115	161.6	81.5	27.3	8.9
Lung	31 231	22 761	99.1	50.5	72.3	34.7
Colon-Rectum	23 216	9 209	73.7	34.0	29.2	11.5

(a) Estimates for men for year 2015 (prostate) and 2018 (lung and colon-rectum)

Localisation	Number of new cases	Number of death	AIR / 100 000 p-y Raw	AIR / 100 000 p-y Standardized	AMR / 100 000 p-y Raw	AMR / 100 000 p-y Standardized
All localizations	177 433	67 817	529.4	274.0	202.3	72.2
Breast	58 459	12 146	174.4	99.9	36.2	14.0
Colon-Rectum	20 120	7 908	60.0	23.9	23.6	6.9
Lung	15 132	10 356	45.1	23.2	30.9	14.0

(b) Estimates for women for year 2018

Table 7.1: Estimated number of new cases and death by cancer and estimates of raw and age-standardized (over the worldwide age distribution) Annual Incidence Rate (AIR) and Annual Mortality Rate (AMR) per 100,000 person-years (p-y) for the three main localizations per sex in France in 2018 (all localizations except prostate) or 2015 for prostate (Defossez et al., 2019). Estimates for prostate cancer is given for 2015 du to high short-term uncertainty for that localization.

Several non-governmental organizations are also actively involved in cancer research funding and patient support, in particular the LNCC[5] (*Ligue Contre le Cancer*), created in 1918, which I especially thank for funding this thesis. Cancer research implies many actors of various disciplines and main research areas include biological mechanisms of carcinogenesis, cancer development and regression, environmental causes and prevention, screening, treatments and patient support.

7.1.2 Cancer genetics

7.1.2.1 Carcinogenesis

Carcinogenesis is a multi-state process conducting to cancer formation. A cell become a tumoral cell by an accumulation of mutations affecting some genes involved directly or indirectly in cell growth and proliferation and conferring a selective advantage to the cell. Via carcinogenesis, a normal cell becomes a malignant tumoral cell. Not all tumors become malignant and lead to a cancer and some of them may even be eliminated by the immune system. A tumor becomes malignant when it invades healthy tissues and threatens their normal functioning. Hanahan and Weinberg (2000) listed six essential characteristics named "*hallmarks of cancer*" to design malignant growth: 1) self sufficiency in growth signal, 2) insensitivity to growth inhibitory signals, 3) escape from apoptosis (physiological, programmed cell death), 4)

[5] `https://www.ligue-cancer.net`

limitless replicative potential, 5) sustained angiogenesis (blood vessel growth) and 6) tissu invasion and metastasis. They added two more hallmarks in (Hanahan and Weinberg, 2011): reprogramming of energy metabolism and escape from immune system. Two main types of genes are involved in cancer:

- *Oncogenes* (named *proto-oncogenes* when not altered) promote cell growth and division. A gain of function in these genes participate in the tumoral process. They act in a dominant manner as a single mutant copy providing a gain of function leads to rescue cells from apoptosis or reduce growth factor dependence leading to loss of growth inhibition. The most well-known oncogenes are HA-RAS and MYC.

- *Tumor-suppressor* genes inhibit cell division and survival and/or are involved in DNA repair. A loss of function in these genes participate in the tumoral process, hence both alleles must be inactivated for them to participate in carcinogenesis. They are divided into two main types (Kinzler and Vogelstein, 1997):

  - *Gatekeeper* genes are involved in cell growth inhibition and death promotion with a dose-dependent function. APC and p53 are two examples of gatekeeper genes.
  - *Caretaker* genes are involved in DNA repair during cell division. Their inactivation leads to genetic instability and finally an increased mutation rate. The most well-known caretaker genes include BRCA1 and BRCA2 as well as genes of the MisMatch Repair (MMR) system (mainly MLH1, MSH2, MSH6, PMS2). The formers (respectively the laters) are mainly associated with breast and ovarian cancer (respectively with colorectal and endometrial cancer). RAD51 also belongs to the class of caretaker genes.

In 2018, the Cancer Gene Census (CGC) within the Catalogue of Somatic Mutations in Cancer[6] (COSMIC) describe the effect of 719 cancer-driving genes (Sondka et al., 2018).

7.1.2.2 Genetic predisposition

Carcinogenesis being a multi-state process involving an accumulation of mutations, it is age dependent and cancer risks increase with age. Most mutations are acquired in life due to environmental and lifestyle factors (tobacco, X-Ray, alimentation, UV, some viruses, etc.). A tremendous literature studies the (positive or negative) effect of various environmental factors on different cancer types. A mutation acquired by a cell during life is said to be somatic. When a mutation affects germ cell, in gonades, it can be transmitted to the next generation. An inherited mutation, called a germline mutation, is carried constitutionally (in all cells) and counts in the accumulation from birth. Therefore, an individual carrying a germline mutation is at

[6] `https://cancer.sanger.ac.uk/cosmic`

higher risk of cancer at younger ages when compared to the general population. An estimated proportion of 5 to 10 % of cancers involves an inherited mutation[7]. Genetic counseling is a discipline which aims at estimating risks of genetic predisposition in order to adapt prevention, screening, surveillance and/or treatments.

7.2 Introduction to the Lynch syndrome

7.2.1 Definition

A microsatellite is the repetition of one to six nucleotide(s) in a DNA sequence. A defective MisMatch Repair (MMR) system leads to tumors characterized by microsatellite instability (MSI). A tumoral DNA containing more than 40% microsatellite variations characterizes a tumor of high MSI frequency, denoted MSI-H. Microsatellite stable (MSS) tumors contain no, or close to no, microsatellite variation. Tumors of MSI low (MSI-L) frequency (less than 40% microsatellite variations) lack clear relevance and they are usually not considered as microsatellite unstable (Sehgal et al., 2014).

The Lynch syndrome (LS), also called somehow confusingly Hereditary Non-Polyposis Colorectal Cancer (HNPCC), is the most frequent genetic predisposition to cancer. It is defined as an inherited mono-allelic mutation, hence a dominant mode of inheritance, in a gene of the MMR system, i.e. mainly MLH1, MSH2, MSH6, PMS2. The EPCAM gene has later been added to LS genes because a constitutional deletion of its 3' end has shown to be associated with LS via epigenetic silencing of MSH2 which is just downstream EPCAM (Ligtenberg et al., 2009). LS confers higher risks of developing colorectal and endometrial tumors earlier in life when compared to the general population. Several other localizations have been added to its spectrum including stomach, ovary, ureter, kidney, small bowel, biliary tract, pancreas, prostate and brain (Vasen et al., 2013). The inclusion of breast cancer into the Lynch spectrum is controversial and most studies reported a much lower penetrance of MMR genes for that localization when compared to other Lynch-associated cancer (Vasen et al., 2013; Dominguez-Valentin et al., 2020b). As seen in the previous section, whereas the mode of inheritance of LS is autosomic dominant, its biological expression is recessive and the inactivation of the gene is necessary to participate in carcinogenesis. Such phenomena can happen either by loss of heterozygosity (deletion of the functional allele), or a pathogenic mutation in the functional allele acquired in life.

The LS was first described in 1895 and published in 1913 by Aldred Scott Warthin who studied a family with numerous cases of colonic, gastric and uterine cancers (Warthin, 1913). It has been characterized in more details in the 70's and the 80's by Lynch and his colleagues (Lynch and Krush, 1971; Lynch et al., 1985a,b) and more and more studied since then.

A biallelic inheritance of pathogenic variants in the same MMR gene is another syndrome called the Constitutional MisMatch Repair Deficiency (CMMRD). The CMMRD syndrome is rare and characterized by pediatric tumors including brain

[7] `https://www.e-cancer.fr`

and colorectal tumors, haematological malignancies and sarcomas often with multiple onsets in childhood (Lavoine et al., 2015).

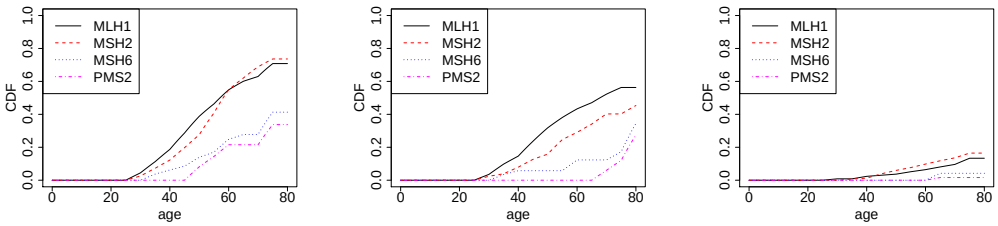
7.2.2 Epidemiology

The LS is involved in 2.8% ([2.1 - 3.8], 95% IC) of colorectal-cancers (CRC) (Hampel et al., 2008), 3.2% ([1.8 - 5.1], 95% IC) of endometrial cancers (EC) (Ryan et al., 2020) and about 15% of MSI CRC (Sourrouille et al., 2013; Mensenkamp et al., 2014). The frequency of heterozygous carriers in the general population is estimated at 0.051% ([0.039 - 0.068], 95% IC) for MLH1, 0.035% ([0.026 - 0.048], 95% IC) for MSH2, 0.132% ([0.089 - 0.196], 95% IC) for MSH6 and 0.140% ([0.094 - 0.208], 95% IC) for PMS2 and an overall LS prevalence at 0.359 ([0.248 - 0.520], 95% IC) in the general population (Win et al., 2017). However LS prevalence is prone to controversies with estimates ranging from one to five fold between studies. Deleterious variants (or alleles) in MSH6 or PMS2 are more frequent in the general population however MLH1 and MSH2 being much more penetrant, mutations in these genes involve more sever histories of Lynch-associated cancers and are more frequently encountered in MSI tumors. De novo mutations are rare and estimates for their prevalence remain unavailable. Note also that MSH2 and MSH6 are in linkage disequilibrium as they are closely located on Chromosome 2 at positions 47,403,067 basepair (bp) to 47,634,501 bp for MSH2 and 47 783 145 bp to 47 806 954 bp for MSH6 (see the National Center for Biotechnology Information website[8]).

The cumulative distribution function (often called cumulative risk in medical genetics) of the time to first diagnosis in main localizations and all localizations mentioned in the previous paragraph is represented in Figure 7.1 and 7.2 respectively for male and female carriers of a single pathogenic mutation. Computations were done under the assumption of piecewise constant hazard functions by steps of 5 years with hazard rates estimated by Dominguez-Valentin et al. (2020b) and referenced by the InSiGHT group[9] (International Society for Gastrointestinal Hereditary Tumors). The probability of developing a colon cancer for carriers of a pathogenic mutation in MLH1 or MSH2 before age 75 is 0.40 to 0.55 for males and 0.40 to 0.50 for females and it goes up to 0.70 for males and 0.80 for females for all Lynch-associated localizations combined. MSH6 is less penetrant for colon cancers but highly penetrant for endometrial cancer with a cumulative risk of 0.40 at 75 years old for MSH6 female carriers. PMS2 is much less penetrant than any of the main three other genes for all localizations.

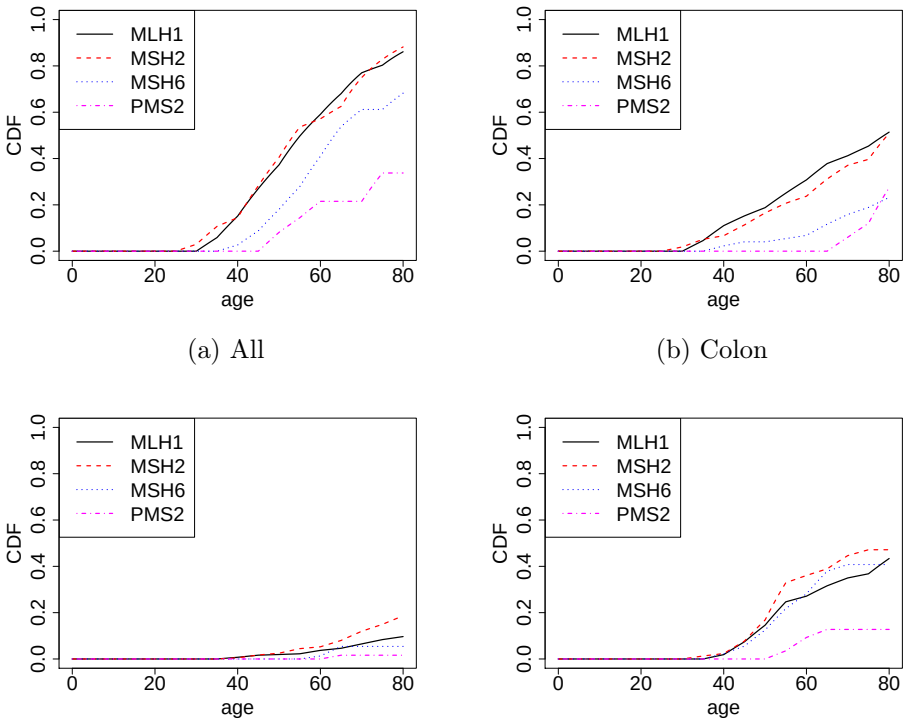
Several covariates participate in the risk of Lynch-associated tumors and MMR deficiency including body mass index, smoking, alcohol and diabetes (Pande et al., 2010; Movahedi et al., 2015; Dashti et al., 2019).

[8] `https://www.ncbi.nlm.nih.gov/genome/gdv/browser`
[9] `https://www.insight-group.org`



(a) All (b) Colon (c) Rectum

Figure 7.1: Cumulative distribution function of time to first diagnosis for all Lynch-associated cancers (on the left) and for main localizations (on the middle and on the right) for males per genotype. An heterozygous carrier of a pathogenic mutation is simply denoted by the corresponding gene. Source for annual incidence rates per genotype and localization: Dominguez-Valentin et al. (2020b).



(a) All (b) Colon

(c) Rectum (d) Endometrium

Figure 7.2: Cumulative distribution function of time to first diagnosis for all Lynch-associated cancers and for main localizations for females per genotype. An heterozygous carrier of a pathogenic mutation is simply denoted by the corresponding gene. Source for annual incidence rates per genotype and localization: Dominguez-Valentin et al. (2020b).

7.3 Lynch syndrome detection and risks assessment

LS detection is of high importance in order to adapt the surveillance of patients and their family members. Clinical guidance for LS patients towards CRC is widely documented (Vasen et al., 2013) but the effects of a surveillance for EC is still uncertain (Crosbie et al., 2019). Currently a colonoscopy every one to two years is recommended for LS patients. Daily aspirin intake significantly reduces risks of CRC (Burn et al., 2011). Hysterectomy and bilateral oophorectomy highly prevents from endometrial and ovarian cancer and is recommended for LS patients at appropriate age no earlier than 35-40 (Crosbie et al., 2019) however pros and cons are discussed for each patient (Schmeler et al., 2006). Efficacy of transvaginal ultrasound and endometrial biopsy is not proved (Vasen et al., 2013). Clinical guidelines are regularly published for guiding practicians and genetic counselors in their evaluation of probabilities of carrying LS for their patients and other family members in order to adapt germline screening prescriptions and/or surveillance.

7.3.1 First criteria

The *Amsterdam criteria I* established by the International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC) were the first published clinical guidelines to offer standardized criteria for LS suspicion (Vasen et al., 1991). They were extended to extra colonic tumors in the *Amsterdam criteria II* (Vasen et al., 1999) and MSI markers were included in the *Bethesda guidelines* (Rodriguez-Bigas et al., 1997). These criteria were combined, revised and published in the *Revised Bethesda Guidelines* by Umar et al. (2004) and are summarized in Table 7.2.

7.3.2 Biological testing and clinical data

7.3.2.1 MSI status

Numerous studies have shown a poor sensitivity and specificity of the Amsterdam and Bethesda criteria (Vasen et al., 2013) but MSI screening for all CRC significantly improves LS detection when associated with Bethesda criteria (Hampel et al., 2008; van Lier et al., 2012; Canard et al., 2012). Furthermore MSI tumors show different responses to chemotherapy (Sargent et al., 2010; Dorard et al., 2011; André et al., 2015) and strong responses to immune checkpoint inhibitors at a metastatic stage rendering MSI status to become a biomarker for inclusion in immunotherapy clinical trials (Le et al., 2015; Overman et al., 2017; Colle et al., 2017; Overman et al., 2018; Cerretelli et al., 2020). Therefore MSI tumors are candidate for personalized medicine. The possible implication of MSI status in EC for providing information on prognosis and treatments has been suggested by several authors (Diaz-Padilla et al., 2013; Talhouk et al., 2015; Stelloo et al., 2016; Sloan et al., 2017). Therefore universal screening of MSI status is now recommended for all CRC and all EC before age 70. There exists no recommendation of MSI detection for other Lynch-associated cancers (Vasen et al., 2013).

Two methods are standardly used to test the MSI status of a tumor: Polymerase Chain Reaction (PCR) and ImmunoHistoChemistry (IHC). PCR-based technics aim

Amsterdam criteria I	1.	At least three relatives with histologically confirmed CRC One of whom is a first degree relative of the other two
	2.	At least two successive generations affected by CRC
	3.	At least one CRC diagnosed before age 50
	4.	Familial adenomatous polyposis excluded
Amsterdam criteria II	1.	At least three relatives with a Lynch-associated cancer One of whom is a first degree relative of the other two
	2.	At least two successive generations affected
	3.	At least one of the Lynch-associated cancers is diagnosed before age 50
	4.	Familial adenomatous polyposis excluded
	5.	Tumors should be verified by pathologic examination
Revised Bethesda	1.	CRC diagnosed before age 50
	2.	Synchronous or metachronous Lynch-associated tumors regardless of age
	3.	MSI CRC diagnosed before age 60
	4.	CRC plus a Lynch-associated cancer in at least one first-degree relative. One of the cancers should be diagnosed before age 50
	5.	CRC diagnosed in two or more first or second-degree relatives with Lynch-associate tumors regardless of age

Table 7.2: Amsterdam criteria I and II and revised Bethesda guidelines (Vasen et al., 1991, 1999; Rodriguez-Bigas et al., 1997).

at comparing alternate-sized microsatellites in tumor and germline DNA. The choice of microsatellite markers is important and a panel of five mononucleotides repeats called Pentaplex panel (Suraweera et al., 2002) is one of the most commonly used (Umar et al., 2004). IHC checks MMR proteins expression in tumor tissue using corresponding antibodies. IHC has the advantages of being rapid and less costly and identifies the affected protein hence the related gene. All four proteins are usually tested together. Note that MMR proteins form heterodimers to build functional complexes MLH1/PMS2 and MSH2/MSH6. A loss of MLH1 (respectively MSH2) protein leads to loss of PMS2 (respectively MSH6) protein. On the contrary a loss of PMS2 or MSH6 does not lead to loss of MLH1 nor MSH2. Therefore, an MLH1/PMS2 (respectively MSH2/MSH6) loss result is in favor of MLH1 (respectively MSH2) deficiency but inconclusive for PMS2 (respectively MSH6), an isolated loss of MSH6 (respectively isolated loss of PMS2) is in favor of MSH6 (respectively PMS2) deficiency. Next Generation Sequencing (NGS) may replace or complement PCR for MSI screening in the future (Salipante et al., 2014; Stadler et al., 2016; Nowak et al., 2017; Hampel et al., 2018).

7.3.2.2 Additional biological testing

The majority of MSI CRC and MSI EC are a consequence of somatic hypermethylations of MLH1 promoter. 80% to 85 % of MSI tumors are not a consequence of LS but a result of biallelic somatic events such as pathogenic mutations, loss of heterozygosity, somatic methylation, etc. in MMR genes (de la Chapelle et al., 2009; Sourrouille et al., 2013; Mensenkamp et al., 2014). The most frequent somatic event is an hypermethylation of MLH1 promoter (Kane et al., 1997; Esteller et al., 1998). BRAF V600E mutation is also a strong predictor against LS and frequently associated with MLH1 promoter hypermethylation in CRC (Parsons et al., 2012). Some studies showed its non-association with MSI EC (Weissman et al., 2012). BRAF screening was used as a surrogate for MLH1 promoter hypermethylation in CRC although direct testing for hypermethylation is now preferred in particular in case of no screened BRAF V600E. Therefore MLH1 promoter hypermethylation testing in CRC and EC and/or BRAF V600E screening in CRC is recommended if MLH1/PMS2 proteins loss is detected by IHC. Note that in some rare cases however MLH1 promoter methylation is inherited via a non-Mendelian epimutation inheritance (Hitchins et al., 2007; Ward et al., 2013) and some rare patients with LS develop MSI tumors with BRAF V600E mutation (Parsons et al., 2012). A cascade of testing for discriminating patients at high versus low probability of LS has been summarized by the INCa and displayed in Figure 7.3 for CRC and Figure 7.4 for EC.

7.3.2.3 Sensitivity and specificity

The sensitivity (respectively specificity) of a test is defined as the probability of a positive (respectively negative) result conditional on a genotype consistent with a positive (respectively negative) result. In this whole chapter and the next one, we define a positive PCR-based MSI test (respectively a positive IHC test towards a

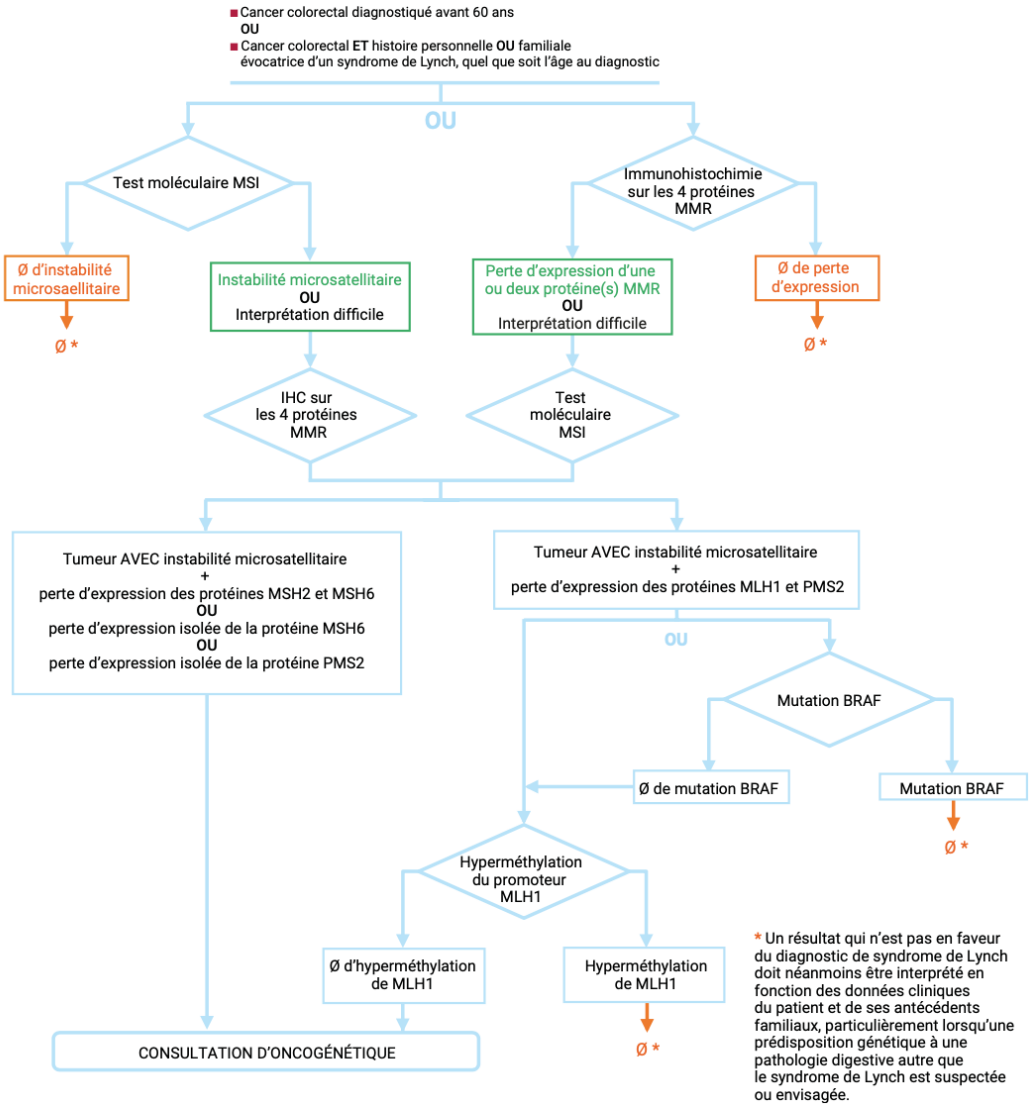


Figure 7.3: Cascade of biological testing on colorectal tumors to discriminate high LS risk patients who should be addressed to genetic counseling versus low LS risk patients. This chart is proposed by the Institut National du Cancer (source: INCa).

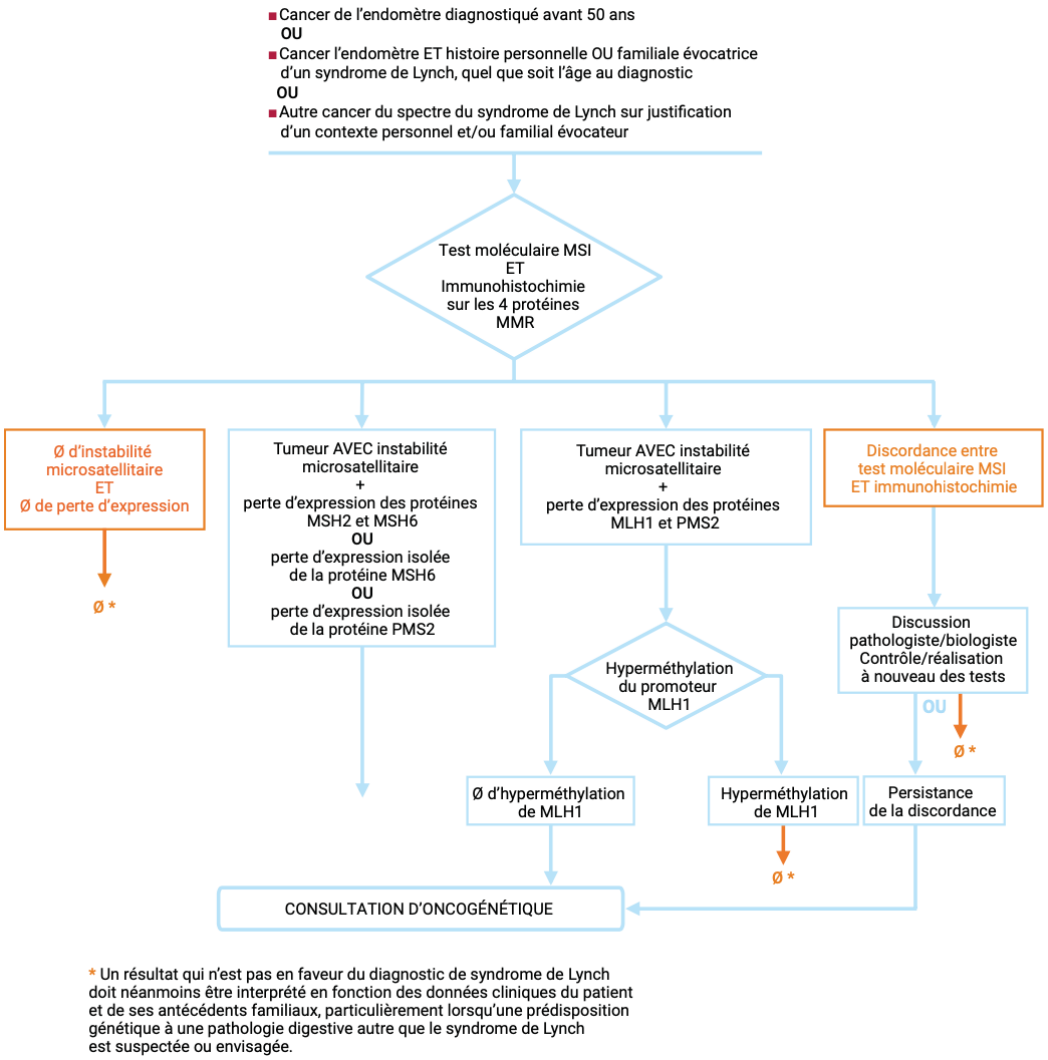


Figure 7.4: Cascade of biological testing on endometrial tumors to discriminate high LS risk patients who should be addressed to genetic counseling versus low LS risk patients. This chart is proposed by the Institut National du Cancer (source: INCa).

	Number of studies	Sensitivity [95% CI]	Number of studies	Specificity [95% CI]
PCR-based MSI				
overall	26	0.940 [0.894 - 0.967]	21	0.754 [0.670 - 0.823]
5 markers	18	0.947 [0.891 - 0.975]	12	0.770 [0.664 - 0.850]
IHC-based MSI				
overall	17	0.900 [0.841 - 0.939]	10	0.810 [0.643 - 0.910]
3 proteins	9	0.907 [0.816 - 0.955]	4	0.851 [0.367 - 0.983]
MLH1 promoter hypermethylation	8	0.820 [0.618 - 0.930]	9	0.960 [0.740 - 0.990]
BRAF PCR	7	0.570 [0.450 - 0.690]	7	0.980 [0.900 - 0.990]

Table 7.3: Sensitivity and specificity of biological tests in colorectal tumors estimated by Assasi et al. (2016) from pooled studies.

protein or a couple of proteins) to be a screened MSI-H tumor DNA (respectively the observed loss of the corresponding protein(s)) and a positive MLH1 promoter hypermethylation test (respectively BRAF V600E mutation screening) to be an observed hypermethylation (respectively a screened V600E mutation in BRAF gene). Therefore, the sensitivity (respectively specificity) of an PCR-based MSI test is defined as the probability of a positive (respectively negative) test conditional on a LS (respectively non-LS) germline genotype and the sensitivity (respectively specificity) of an IHC-based MSI test is defined as the probability of a positive (respectively negative) test conditional on a pathogenic germline mutation (respectively no pathogenic germline mutation) in the corresponding gene. Moreover the sensitivity (respectively specificity) of MLH1 promoter hypermethylation and BRAF V600E tests are defined as the probability of a positive (respectively negative) test conditional on no pathogenic germline mutation in MLH1 (respectively a pathogenic germline mutation in MLH1).

The sensitivity and specificity of biological tests in CRC have been widely studied. Svrcek et al. (2019) compared PCR and IHC technics in CRC and concluded that both methods are equally valid. Assasi et al. (2016) conducted a meta-analysis for estimating the sensitivity and specificity of the aforementioned tests in CRC from pooled studies with detailed exclusion criteria. Estimates for PCR-based MSI were computed from overall pooled studies and pooled studies restricted to five markers. Results reported by the authors are summarized in Table 7.3.

Estimates of the sensitivity and specificity of MSI-PCR and IHC tests for EC are much scarcer in the literature. Some studies report good concordance between both methods (McConechy et al., 2015; Stelloo et al., 2017; Raffone et al., 2020) but only few of those germline screened enough patients for sensitivity and specificity estimations. In Crosbie et al. (2019) the authors list some of them with reported sensitivity (respectively specificity) for IHC ranging from 86 to 100% (respectively 48 to 67 %) and sensitivity (respectively specificity) for MSI-PCR ranging from 77 to 100% (respectively 38 to 81 %). However (Ryan et al., 2020) reports a much higher sensitivity of IHC versus PCR in EC (100% versus 56%) and equal specificity

(97.5%). EC cohorts for parameter estimations are usually small and results should be interpreted with caution. IHC is usually advocate in particular because MSI-PCR based triage may miss MSH6 pathogenic variants. Indeed MSH6 is highly penetrant for EC and less often associated with MSI-H status (Wu et al., 1999; de Leeuw et al., 2000). Combining both methods for EC has been proven useful (Goodfellow et al., 2015).

Other clinical data Some clinical data are additional predictors for LS suspicion versus sporadic tumors. In particular, several authors have shown since the 90' that colon cancer linked to LS more frequently occur in the proximal colon (Lynch and Smyrk, 1996; Järvinen et al., 2000). More recently, the association between histopathological profile of ovarian carcinoma and LS carrier status is studied and some authors suggest that ovarian carcinoma in LS patients are more frequently of non-serous type (Ketabi et al., 2011; Pal et al., 2012; Chui et al., 2014; Nakamura et al., 2014; Rambau et al., 2016; Crosbie et al., 2020). However precise estimates of the sensitivity and specificity of that latter clinical profile remain unavailable.

7.3.3 Mathematical models

With an increasing number of family history data, clinical and biological testing results, mathematical models for computing continuous probabilities of genetic predisposition are increasingly useful. There exists two main types of mathematical models: logistic regressions and pedigree-based models.

Logistic regression. Logistic regression based models in the context of Lynch-associated cancers are MMRPredict (Barnetson et al., 2006) and PREMM models. PREMM is the only model regularly updated with a last version called PREMM_5 released and evaluated in 2017 (Kastrinos et al., 2017). Logistic regression based models are individual oriented and compute probabilities of carrying deleterious mutations for a targeted individual (called proband).

PREMM_5 was developed from 18,734 individuals who underwent germline screening for the five genes MLH1, MSH2, MSH6, PMS2 and EPCAM. Individuals carrying more than one mutation in one MMR gene were excluded and carriers of MSH2 and EPCAM mutations were pooled leading to five considered genotypes: non-carrier, carrier MLH1, carrier MSH2 or EPCAM, carrier MSH6 and carrier PMS2. A polytomous logistic regression was performed in order to estimate associations (as odd ratios) between chosen covariates and genotypes. Retained covariates are listed in Appendix Table A1 in (Kastrinos et al., 2017) and include sex and current age of the proband, history of cancer with age at diagnoses for CRC and EC for the proband and first and second-degree relatives as well as partial data for other Lynch-associated cancer. Multiple cancers affecting a single individual are considered but the age at diagnosis of the first cancer is solely retained. Family history is restricted to affected first and second degree relatives. The model was primary developed before universal screening for MSI status recommendations. Hence, despite the increasing availability of biological testing data, none of these data nor clinical profiles are included.

Estimated regression coefficients and equation is available in the Appendix section of (Kastrinos et al., 2017) from which one can derive a posterior probability of carrying a pathogenic mutation per gene or an overall posterior probability of carrying LS for the proband. In its user-friendly online version[10], PREMM_5 computes an overall posterior LS probability.

Pedigree-based models. Pedigree-based models were introduced in Chapter 3 and in particular in Section 3.2. They are Bayesian Networks (BNs) and model the structure dependency between genotypes of family members and between genotypes and phenotypes. Hence they present the main advantage of taking into account the entire structure dependency between individuals and the entire dataset composed of family history of cancer (for both affected and unaffected individuals whatever his relative degree with the proband) and covariates. They also are generating, explicative and can be used for simulating families and computing risks for any family member. Their principale disadvantage is their high computational cost for exact inference which can however be dropped with the message-passing algorithm developed in Chapter 1.

The only current pedigree-based model for Lynch-associated cancers is MMRPro (Chen et al., 2006) which computes posterior probabilities of carrying a deleterious mutation in MLH1, MSH2 and/or MSH6 for the proband or any family member conditional on a family history of CRC and EC and biological testing reduced to MSI status by PCR-based of IHC-based technic. It was developed in 2006 with parameters estimated from data extracted from the literature. Classical assumptions in pedigree based models are made, i.e. Assumption A1 (genes are biallelic), A2 (genotypes of founders are in Hardy-Weinberg equilibrium), A3 (alleles for a given gene segregate independently) and A4 (no genomic imprinting). Furthermore the authors assume that genes segregate independently, hence they ignore the non-independent segregation of MSH2 and MSH6 in linkage disequilibrium. However that assumption induces close to no bias in the computation of posterior probabilities of carrying LS. A personal history of cancer is assumed to be independent of clinical data conditional on the genotype of the individual. Their survival model is a competing risk model as the one represented in Figure 3.5 with two diseases and three states: State 0 stands for "Unaffected", State 1 for "Diagnosed with CRC" and State 2 for "Diagnosed with EC" and a baseline Weibull distribution. Multiple cancers affecting the same individual are not supported and the first diagnosed cancer is the only one taken into account.

Each allele take value 0 if non-pathogenic and 1 otherwise. The genotype associated with gene $g \in \{\text{MLH1}, \text{MSH2}, \text{MSH6}\}$ for an individual $i \in \{1, \dots, n\}$ in a family of n members is assumed to be equal to the sum of its (paternal and maternal) component alleles and denoted X_i^g. All genotypic combinations are assumed to be viable, therefore the set of values for individual genotypes, all genes combined, $X_i = \left(X_i^{\text{MLH1}}, X_i^{\text{MSH2}}, X_i^{\text{MSH6}}\right)$ is $\{0, 1, 2\}^3$. Transition intensities are sex and genotype dependent. Conditional on a non-carrier genotype, they are assumed to be equal those in the general population and extracted from incidence

[10] `https://premm.dfci.harvard.edu`

data per sex and localization estimated by the SEER[11] (Surveillance, Epidemiology and End Results program) registry. Conditional on a genotype composed of at most one pathogenic allele $X_i \in \{(1,0,0),(0,1,0),(0,0,1)\}$, transition intensities are estimated from a meta-analysis of five references as detailed in (Chen et al., 2006). The chosen form of the hazard and the method used for parameter estimation is not mentioned by the authors. Adequately with LS mode of inheritance, the authors assume a dominant mode of inheritance, hence a dominant hazard model between non-carrier and carrier genotypes. Transition intensities conditional on a genotype composed of more than one pathogenic mutation $X_i \in \{0,1,2\}^3 \setminus \{(0,0,0),(1,0,0),(0,1,0),(0,0,1)\}$, are obtained under the assumption of an additive effect of pathogenic alleles on the hazard. Consequently, denoting by $\lambda_k^{s,x}$, the transition intensity from State 0 to State $k \in \{1,2\}$ conditional on sex $s \in \{1,2\}$ and genotype $x = (x^{\text{MLH1}}, x^{\text{MSH2}}, x^{\text{MSH6}}) \in \{0,1,2\}^3 \setminus (0,0,0)$, the authors assume that

$$\lambda_k^{s,x} = x^{\text{MLH1}} \times \lambda_k^{s,(1,0,0)} + x^{\text{MSH2}} \times \lambda_k^{s,(0,1,0)} + x^{\text{MSH6}} \times \lambda_k^{s,(0,0,1)}.$$

Allele frequencies in the general population are computed from estimates of allele frequencies among cases and incidences per localization in the general population and among carriers extracted from the literature. Furthermore, the sensitivity and specificity of biological testing are computed from meta-analyses of pooled literature as detailed in (Chen et al., 2006). Finally the method developed in Section 3.2 is applied to compute posterior carrier probabilities and classical formulae to compute CPDs of the phenotypic component in the competing risk model chosen by the authors are recalled in Equations (3.3) and (3.4). MMRpro is implemented in the BayesMendel R package available for non-clinical research upon request to the BayesMendel Lab[12].

Validation. PREMM_5 was validated on a clinical-based cohort of 1,058 patients with CRC and distinguished carrier from non carriers with and AUC of 0.81 ($[0.75 - 0.92]$, 95% IC) for all genes pooled and showed better results for discriminating carriers of highly penetrant genes with AUC of 0.89 ([0.87 - 0.91] 95% CI) for MLH1, 0.84 ([0.82 - 0.86] 95% IC) for MSH2/EPCAM and 0.76 ([0.73 - 0.79] 95% IC) for MSH6 but 0.64 ([0.60 - 0.68] 95% IC) for PMS2. Its sensitivity (respectively specificity) was estimated at 0.894 (respectively 0.492) with a threshold at 2.5% and 0.721 (respectively 0.751) with a threshold at 5%. Kastrinos et al. (2017) compared their earlier version $\text{PREMM}_{1,2,6}$ and their most recent version PREMM_5 and concluded to better performances of PREMM_5 over $\text{PREMM}_{1,2,6}$ with threshold probability of carrying a LS at 2.5%, threshold recommended by the authors.

MMRpro was validated on a small clinical-based cohort of 279 individuals from 226 families in the United States, Canada, and Australia. It discriminated LS carriers versus non-carriers with an AUC of 0.83 ([0.78 - 0.88] 95% CI).

Predictive performances of MMRpredict, MMRPro and the version $\text{PREMM}_{1,2,6}$ of PREMM in their probability of LS assessments over population-based and clinical-

[11] `https://seer.cancer.gov/registries/`

[12] `https://projects.iq.harvard.edu/bayesmendel`

based samples were compared by Win et al. (2013) who concluded to similar performances of the three models with AUC ranging from 0.80 ([0.72 - 0.88] 95% CI) to 0.84 ([0.81 - 0.88] 95% CI).

Implementation and current guidelines. In the framework of breast/ovarian cancer, numerous mathematical models regularly updated and of both types exist for assessing probabilities of genetic predisposition and disease risks. Main logistic regression based models include the BCRAT[13] (Breast Cancer Risk Assessment Tool) model, formerly named the Gail model (Gail et al., 1989; Gail, 2015), the CARE (Women's Contraceptive and Reproductive Experiences) (Gail et al., 2007) and AABCS (Asian American Breast Cancer Study) (Matsuno et al., 2011) models which are respectively extensions of the Gail model to African American women and Asian/Pacific Islander American women. Another logistic regression based models developed on a large cohort and accounting for other associated risks factors such as BMI, menopausal hormones and alcohol consumption was developed by Pfeiffer et al. (2013).

Main pedigree-based models in the framework of breast/ovarian cancer include BOADICEA (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) (Antoniou et al., 2002, 2004, 2008; Lee et al., 2014, 2019), IBIS (Tyrer et al., 2004; Brentnall et al., 2019), the Claus model (Claus et al., 1991; Easton et al., 1993; Claus et al., 1994) and BRCApro (Berry et al., 1997; Chen and Parmigiani, 2007; Mazzola et al., 2014, 2015). Their relative performances are compared in several studies (de Pauw, 2012; Cintolo-Gonzalez et al., 2017; McCarthy et al., 2020).

In contrast, mathematical models for LS and Lynch-associated cancers risks assessments are much scarcer, mainly restricted to PREMM_5 and MMRpro. They also are poorly implemented in daily clinical practice (Lynch et al., 2007) and absent from current guidelines (Vasen et al., 2013; Syngal et al., 2015) except in the National Comprehensive Cancer Network[14] (NCCN) which advocates the use of PREMM_5 with a threshold at 5% (Boland et al., 2018). The rare and inadequate offer could be one of the main reasons for explaining the underuse of mathematical models in LS risk assessment.

MMRpro was developed in 2006 and, to the best of our knowledge, not updated since then. It showed good predictive performances on its (small) validation cohort. Extra-colonic, extra-endometrial Lynch-associated cancers are ignored in calculation despite their inclusion in the Amsterdam II (Vasen et al., 1999) and Bethesda guidelines (Rodriguez-Bigas et al., 1997). Biological testing results are included but restricted to MSI status via PCR-based or IHC and pooled. MLH1 promoter hypermethylation tests and BRAF V600E screening are excluded. Moreover the chosen survival model is not designed for handling multiple cancers affecting a single individual which are however quite frequently encountered in LS patients. Let us finally note that the set of genotypes composed of all possible combinations of alleles seems too wide and, as briefly mentioned in Section 7.2.1, cancer types and incidences in

[13] `https://bcrisktool.cancer.gov`

[14] `https://www.nccn.org`

CMMRD syndrome carriers (two pathogenic mutations in the same gene) are not concordant with an additive hazard model for that set. However that latter assumption induces close to no bias in the computation of posterior probabilities of carrying LS.

$PREMM_5$ showed good predictive performances on its large validation cohort and is easy to use via limited input data and an online implementation. As for any logistic regression based model, family history of cancer is solely partly considered. However despite data restriction, that last point presents the important advantage of simplifying parameter estimations as well as easing their use in daily clinical practice. However $PREMM_5$ may suffer from the fact that no biological testing result is considered. Universal MSI status screening now recommended for all CRC and EC before age 70 renders such data increasingly available (Vos et al., 2020). The high sensitivity and specificity of such tests render them important input data for the computation of posterior probabilities of LS.

Our goal. We believe that mathematical models based on family histories of cancer and clinical and biological data could be an important complementary tool for a primary objective and standardized computation of posterior probabilities of carrying LS for a proband and other family members if accompanied by the expertise of a clinician or a genetic counselor. With an increasing availability of biological testing results such model should be designed for integrating data along their availability to assess risks at each stage of the process and guide clinicians and counselors in subsequent investigations and patients surveillance (Benusiglio et al., 2020).

Furthermore a patient with an MSI CRC or MSI EC but no evidence of biallelic somatic MMR inactivation nor pathogenic variant sequenced in an MMR gene is considered as Lynch-Like (LL) and often undergo same surveillance as a confirmed LS patient with a germline sequenced pathogenic variant. Indeed no screened germline mutation could be the result of lack of sensitivity of current methods or pathogenic mutations in unknown genes. LL conclusion could however also be due to lack of sensitivity of somatic biological testing. Rodríguez-Soler et al. (2013) conducted a large study in order to determine risks for LL patients and concluded that risks of CRC were significantly lower in LL patients than confirmed LS patients but higher than risks in families with sporadic CRC. Therefore assessing a continuous probability of LS for LL patients based on their family history and testing results could be an important information for clinicians for LL patients surveillance.

Moreover the rapid and recent developments of NGS technics lead to a growing number of germline sequenced Variants of Unknown Significance (VUS) whose biological neutrality or pathogenicity is insufficiently known and leaves patients and clinicians with uncertain conclusions. The InSIGHT database[15] references MMR variants. Variants are classified on a five-tiered categorical scale according to their probability of pathogenicity evaluated by multidisciplinary committee as detailed by Thompson et al. (2014) using clinical and functional data available as well as multifactorial likelihood models such as (Easton et al., 2007) or (Goldgar et al., 2008).

[15] `http://www.insight-database.org/genes`

Variants are classified as “pathogenic” (Class 5), “likely pathogenic” (Class 4), “uncertain” (Class 3), “likely not pathogenic” (Class 2) or “not pathogenic” (Class 1). Unfortunately class 3 variants are the most common ones awaiting for more data to refine their classification.

In that context, the purpose of the next chapter is the development of a statistical model called *LynchRisk* to offer a tool for clinicians and genetic counselors in their assessment of posterior probabilities of carrying LS and developing Lynch-associated cancers. We assume that the most exhaustive exploration of Lynch-associated cancer histories in a family is important and therefore, LynchRisk is a pedigree-based model and takes into account the detailed and entirely reported family history of cancers, including affected and unaffected individuals and computes risks for any family member. Therefore it is also a tool for selecting individuals at higher risk within a whole family as well as families at risk.

We aim at overcoming limitations of existing models mentioned in the previous paragraph with parameters and/or exhaustive statistics extracted from recent literature. Multi-state survival models with competing risks and transitions between cancer types are implemented in order to take into account multiple localizations and allow for considering multiple diagnoses per individual in a rigorous manner. Linkage disequilibrium between MSH2 and MSH6 can be taken into account at an additional computational time complexity. Along with the detailed family history, LynchRisk integrates biological testing results composed of MSI screening, IHC testing results per protein or protein dimer, MLH1 promoter hypermethylation and BRAF V600E screening whenever available and thus, it mimics the decision process of clinicians in a reproductible manner. LynchRisk could also be associated with co-segregation analyses as a complementary tool for variants classification. This constitute a perspective of the work.

Chapter 8

LynchRisk model

Sommaire

This chapter is devoted to the development of a new pedigree-based model, named *LynchRisk*, in the framework of the Lynch syndrome. It constitutes the last contribution of the thesis. LynchRisk includes the main four MMR genes (MLH1, MSH2, MSH6, PMS2) and its principal objective is to identify individuals and families at high probabilities of LS. LynchRisk computes the probability of carrying a LS (per

gene or overall LS) for any family member as well as the distribution of the number of carriers in a family and individual probabilities conditional on that number. These probabilities are computed conditional on clinical and biological testing results and detailed family history of cancer in colon (CC), rectum (RC), endometrium (EC), ovary (OC), upper gastrointestinal tract which aggregates stomach, small bowel, bill duct, gall bladder and pancreas (GIC) and urinary tract which aggregates ureter, kidney and urinary bladder (UC). The choice of these localizations was determined by parameter estimation detailed in Section 8.3.4. Additionally, LynchRisk computes individual and familial future cancer risks. A summarized list of input data and output is given in Table 8.1. Except parental relationships, any input can be omitted, in particular of course, clinical and biological testing results when not applicable. The family history of cancer should be reported as exhaustively as possible.

All notions seen in component chapters of Part I are exploited in this chapter. This chapter is organized as follows: we start in Section 8.1 with an introduction to component variables of the model. In Section 8.2 we recall notions seen in Section 3.2, in terms of probabilistic relationships between component variables leading to their modeling into a Bayesian network (BN). We also detail component local probability distributions along with parameters of the model and main assumptions. Parameters of the model are direct extractions from recent literature except transition intensities for time-to-event data for which estimates reported in the literature are insufficiently detailed. We explain in Section 8.3 the method we used for estimating transition intensities from available data and estimates reported in the literature along with additional assumptions that become unavoidable. Section 8.4 is an application section in which we propose a selection of computed quantities of interest, mainly individual and familial posterior carrier risks and future disease risks, over various simulated datasets. We show the qualities of the model along with its drawbacks in comparison with PREMM_5 and MMRpro. Finally after a brief summarized view of the context LynchRisk is built in, we step by step detail future perspectives in Section 8.5.

8.1 General expression and component variables

LynchRisk is built under common assumptions in genetic analysis (Section 3.2), i.e. A1 (biallelic genes), A2 (Hardy-Weinberg equilibrium for genotypes of founders), A3 (independent segregation of alleles per gene) and A4 (no genomic imprinting). We consider a family of n members and we assume that

$$\mathbb{P}(X, Y|S, \theta) = \mathbb{P}(X, Z, B|S; \theta) = \mathbb{P}(X|q)\mathbb{P}(Z|X, S; \alpha)\mathbb{P}(B|X; \beta) \tag{8.1}$$

where $\theta = \{q, \alpha, \beta\}$ is a parameter including allele frequencies in the General Population (GP), incidences of diseases per sex, localization and genotype, sensitivity and specificity of clinical and biological tests, $X = \{X_i\}_{i=1,\dots,n}$ is the set of (latent) genotypes in the family, $Y = \{Y_i\}_{i=1,\dots,n}$ is the set of (usually observed) phenotypes composed of a set of (usually size n) time-to-event data $Z = \{Z_i\}_{i=1,\dots,n}$ and a (usually sparse or empty) set of biological testing results $B = \{B_i\}_{i\in\mathcal{B}}$ where $\mathcal{B} \subseteq \{1, \dots, n\}$ is the set of individuals for whom testing results are available. $S = (S_i)_{i=1,\dots,n}$ is

Input

1. Pedigree
   - structure of the pedigree (exhaustive parental relationships).
   - Sex of each individual.
2. Family history of Lynch-associated cancers up to two diagnoses per individual.
   - Status (unaffected or diagnosed) towards cancers in colon, rectum, endometrium, ovary, upper gastro intestinal tract, urinary tract.
   - Age at diagnosis for affected individuals.
   - Current age or age at death or last news for unaffected (optional for affected) individuals.
3. Clinical and biological tests results (if available)
   - MSI status, IHC towards MMR proteins, MLH1 promoter hypermethylation on colorectal and endometrial tumors, BRAF V600E screening on colorectal tumors.
   - Colon tumor location (proximal or distal).
4. Other clinical data
   - Hysterectomy and age at surgery.
   - Bilateral saplingo-oophorectomy and age at surgery.

Output

1. Individual probability of carrying LS (overall and per gene MLH1, MSH2, MSH6, PMS2) for any family member.
2. Future cancer risks for unaffected individuals and individuals diagnosed with at most one cancer.
3. Distribution of the number of carriers in the family.
4. Individual probability of carrying LS conditional on that number.

Table 8.1: Input/Output in LynchRisk.

the vector of sex of the individuals such that S_i takes value 1 (respectively 2) if individual $i \in \{1, \dots, n\}$ is a male (respectively a female).

Genotypes We denote by $\mathcal{G} = \{G^1, G^2, G^3, G^4\}$ be the set of genes such that G^1 (respectively G^2, G^3 and G^4) stands for MLH1 (respectively MSH2, MSH6 and PSM2), by X_i^g, for $g \in \{1, \dots, 4\}$, the genotype carried by individual $i \in \{1, \dots, n\}$ for gene G^g and by $X_i = (X_i^g)_{g=1,\dots,4}$, the (overall) genotype carried by individual i, i.e. the vector of genotypes per gene. Under A1, each gene is assumed to be biallelic and its alleles take value 0 if non-pathogenic and 1 otherwise. Under A4 we assume that the value taken by X_i^g is equal to the sum of its component (paternal and maternal) alleles. Hence, with no other assumption we have $X_i^g \in \{0, 1, 2\}$, $X_i \in \{0, 1, 2\}^4$ and $X \in \{0, 1, 2\}^{4 \times n}$.

The CMMRD syndrome (see Section 7.2.1) is not taken into account in the first version of LynchRisk. Let us recall that the CMMRD syndrome, defined as biallelic mutations in a MMR gene (two mutations in the same gene) is a cause of (often multiple) brain and gastrointestinal cancers as well as haematological malignancies in childhood (Buecher et al., 2019). As LynchRisk models the transmission of latent alleles between family members, monozygous mutated genotypes should be taken into account. However estimates of incidences per CMMRD-associated cancers are unavailable du to insufficient data. The non-inclusion of CMMRD genotypes in the first version of LynchRisk (as if that syndrome were lethal) seems reasonable. Indeed this syndrome is rare and associated with an early sombre prognosis, thus CMMRD patients rarely have descendants. However its inclusion is important and constitute a perspective for future versions of LynchRisk firstly as a binary variable, secondly with estimated incidences per CMMRD-associated cancer type if more data become available. Therefore, in the first version of LynchRisk, for all $i \in \{1, \dots, n\}$, for all $g \in \{1, \dots, 4\}$, the state space of X_i^g is reduced to $\{0, 1\}$.

To the best of our knowledge, carriers of three or more pathogenic mutations in different MMR genes have never been reported in the literature. This could be du to an extreme rarity of these genotypes or their lethality. We make the assumption of their lethality leading to reducing the state space of each individual genotype X_i to $\mathcal{X} = \{(0000), (1000), (0100), (0010), (0001), (1100), (1010), (1001), (0110), (0101), (0011)\}$. Note that there is no restriction on the state space of genotypes in MMRpro, i.e. any allelic combination is left possible such that each individual genotype takes its values in $\{0, 1, 2\}^3$ (to the power three as PMS2 is not considered in MMRpro). On the contrary, the authors of PREMM_5 assume a maximum of one mutation in one MMR gene, such that the state space of each genotype is reduced to $\mathcal{X}^* = \{(0000), (1000), (0100), (0010), (0001)\}$. Pathogenic alleles being rare in the GP, setting the state space of each X_i to $\{0, 1, 2\}^4$, $\mathcal{X}$ or $\mathcal{X}^*$ has only little consequence on posterior carrier risks computations.

Reducing the state space of component variables of a Bayesian network is defined as entering an *evidence* (see Section 1.1.1.2). Hence, state space reductions, for all $i \in \{1 \dots, n\}$ and $g \in \{1, \dots, 4\}$, $X_i^g \in \{0, 1\}$ and $X_i \in \mathcal{X}$ are evidences for variables of the BN defined Equation (8.1) and will bot be considered before Section 8.4.

Time-to-event variable Each time-to-event variable Z_i is a personal history of Lynch-associated cancer in the set of diseases $\mathcal{D} = \{\text{CC}, \text{RC}, \text{EC}, \text{OC}, \text{GIC}, \text{UC}\}$. We assume that an individual can encounter at most two cancers before age of last news and therefore, the first two cancers (relapses and metastasis excluded) are solely taken into account. The multi-state survival model represented in Figure 8.1 with sex and genotype dependent transition intensities used to compute the individual contribution of each time-to-event data to the likelihood will be detailed in Sections 8.2.2.3 and 8.3.4.

The family history of Lynch-associated cancer should be fulfilled as exhaustively as possible with age at first and second (if applicable) diagnosis and cancer type, age at last news or death for healthy individuals and individuals diagnosed with one cancer. Uncertain phenotypes are considered such that 1) an uncertain cancer type (for instance EC or OC, etc.), 2) a time interval (such that diagnosed with EC between age 40 and 50, Dead before 70, etc.) 3) an uncertain status (such that dead at 60 with no data on previous history). Any unreported individual will be added if he/she has offsprings with reported phenotypes. His/her phenotype is assumed to be missing at random and removed (typical case of an abandon). An individual with missing phenotype is removed if he/she has no descendant with reported phenotypic data and he/she falls into one of the following situation: 1) his/her sex is unknown or 2) his/her sex is known and his/her blood relative with the proband is strictly above third degree. Otherwise he/she is assumed to be healthy and age at censoring is imputed according to ages of other family members. Indeed we assume that most unreported phenotypes falling into the latter situation are likely to be unaffected because of memory bias.

Clinical and biological testing. Along with the family history of cancer, several biological data are strong predictors of LS-linked or sporadic cancer, as previously detailed in Section 7.3.2. In particular, universal MSI screening is now recommended for all MSI CRC and EC diagnosed before age 70 (Vasen et al., 2013) either by PCR or IHC or a combination of both methods. Additional biological testing in MSI tumors are recommended for discriminating between a sporadic and a germline origin of MLH1 deficiency such as MLH1 promoter hypermethylation in colorectal cancers (CRC) and EC and/or BRAF V600E mutation screening in CRC. A cascade of biological testing is summarized on the INCA's website (Figures 7.3 and 7.4). Some clinical data constitute other LS predictors such as colon cancer localization and histopathological profile of ovarian carcinoma (see Section 7.3.2). In the absence of reliable estimates for sensitivity and specificity of histopathological profile of ovarian carcinoma, that latter clinical variable is absent in the first version of LynchRisk.

Let $\mathcal{T} \subseteq \{1, \ldots, n\}$ be the set of individuals who received biological testing results on a CRC and/or an EC, for all $i \in \mathcal{T}$, we add a variable $B_i \subseteq \mathcal{B}_i = \{\text{MSI}_i, \text{IHC.H1.S2}_i, \text{IHC.H2.H6}_i, \text{IHC.iso.H6}_i, \text{IHC.iso.S2}_i \text{Hyper}_i, \text{BRAF}_i, \text{LOC}_i, \widetilde{\text{MSI}}_i, \widetilde{\text{IHC.H1.S2}}_i, \widetilde{\text{IHC.H2.H6}}_i, \widetilde{\text{IHC.iso.H6}}_i, \widetilde{\text{IHC.iso.S2}}_i, \widetilde{\text{Hyper}}_i, \widetilde{\text{BRAF}}_i, \widetilde{\text{LOC}}_i\}$ where no tilde symbol (respectively a tilde symbol) stands for first (respectively second) disease and the names of the variables speak for themselves. Note that $\mathcal{T}$ is usually empty or reduced to the proband and $\mathcal{B}_i$ is usually sparse. A positive result for colon

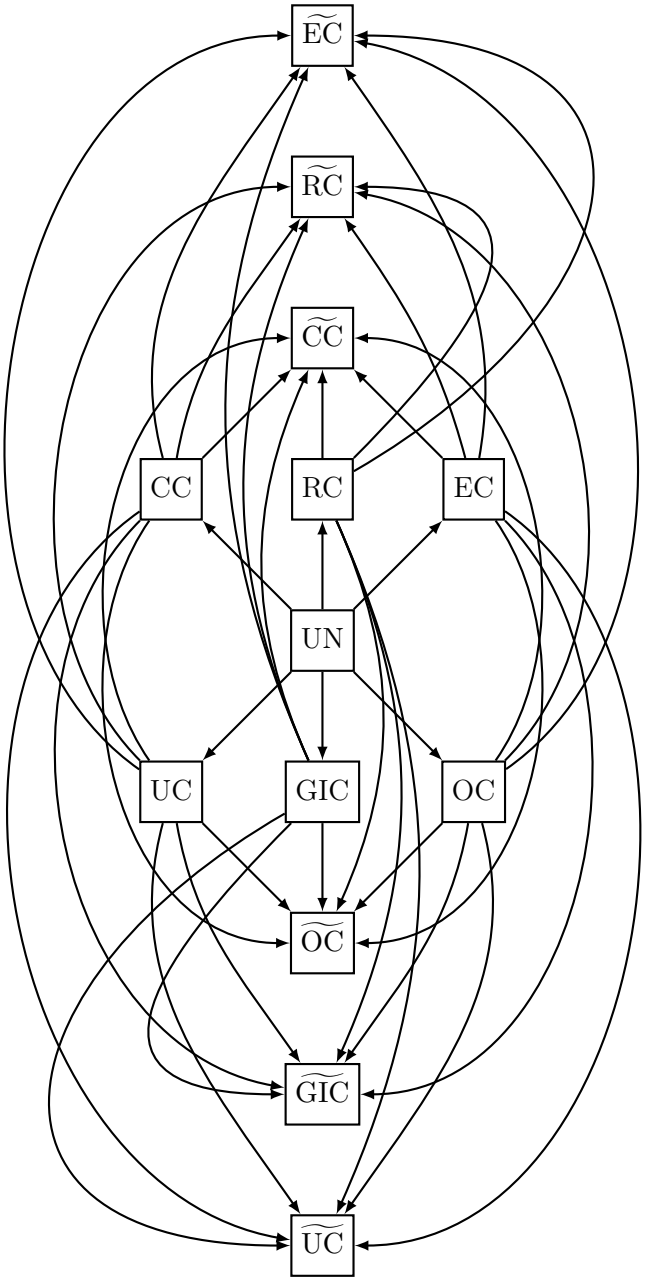


Figure 8.1: Multi-state survival model in LynchRisk where a tilde symbol denotes a subsequent cancer.

localization will be defined as proximal and a positive (respectively negative) result for other variables are defined as in Section 7.3.2.3. Each variables in the set B_i take value 1 for a positive result and 0 otherwise.

8.2 Implementation into a Bayesian network

A detailed explanation for building a BN in pedigree analysis under assumptions A1, A2, A3 and A4 is proposed in Section 3.2. In this section we decompose Equation (8.1) into a genotype, an allele and a selector BN, i.e. we adapt Equations (3.2), (3.5) and (3.6) introduced in Section 3.2 to our context and detail component CPDs and parameter in the second part of the section. LynchRisk's parameters are extracted from estimates available in recent literature and adjusted if needed.

We denote by $\mathcal{F} \subseteq \{1, \dots, n\}$ (respectively $\overline{\mathcal{F}} = \{1, \dots, n\} \setminus \mathcal{F}$), the set of founders (respectively non-founders) in the family and by $\mathcal{T} \subseteq \{1, \dots, n\}$, the set of individuals diagnosed with a CRC and/or an EC who received clinical and/or biological testing results on their tumor(s).

8.2.1 Graph structure

Genotype BN. As mentioned in Section 3.2 a genotype BN is the most intuitive but usually not the most advisable BN, especially when several genes are considered, as the cardinality of each overall genotype is exponential in the number of genes. However it allows for a clear visual interpretation. LynchRisk's genotype BN is written as:

$$\mathbb{P}\left(X, Y | S, \theta\right) = \prod_{i=1}^{n} \underbrace{\mathbb{P}\left(X_i | X_{\mathrm{p}(i)}, X_{\mathrm{m}(i)}; q\right)}_{\text{genotypic component}} \underbrace{\mathbb{P}\left(Y_i | X_i, S_i; \alpha, \beta\right)}_{\text{phenotypic component}} \tag{8.2}$$

where for $i \in \overline{\mathcal{F}}$, $\mathrm{p}(i)$ (respectively $\mathrm{m}(i)$) is the index of the father (respectively mother) of individual i and for $i \in \mathcal{F}$, $\mathrm{p}(i) = \mathrm{m}(i) = 0$ and $X_0 = \emptyset$ by convention. The parameter θ is given by $\theta = \{q, \alpha, \beta\}$ where $q = (q_1, q_2, q_3, q_4)$ is the vector of pathogenic allele frequencies such that, for $g \in \{1, \dots, 4\}$, q_g denotes the frequency of pathogenic alleles in gene G^g in the general population (GP), β is the set of sensitivity and specificity of clinical and biological tests and transition intensities of LynchRisk's survival model represented in Figure 8.1 are parametrized by α. We assume that, for all $i \in \{1, \dots, n\}$,

$$\mathbb{P}\left(Y_i | X_i, S_i; \alpha, \beta\right) = \mathbb{P}\left(Z_i | X_i, S_i; \alpha\right) \mathbb{P}\left(B_i | X_i; \beta\right)^{\mathbb{1}_{\{i \in \mathcal{T}\}}}. \tag{8.3}$$

Allele and selector BN. LynchRisk is implemented into an allele and a selector BN which respectively ignores and takes into account the linkage disequilibrium between MSH2 and MSH6, both closely located on chromosome 2. Taking into account linkage disequilibrium between genes leads to a steep increase of the treewidth of resulting factor graphs, hence a steep increase of computational complexity, even with selector BNs and good elimination orderings. Whereas linkage disequilibrium

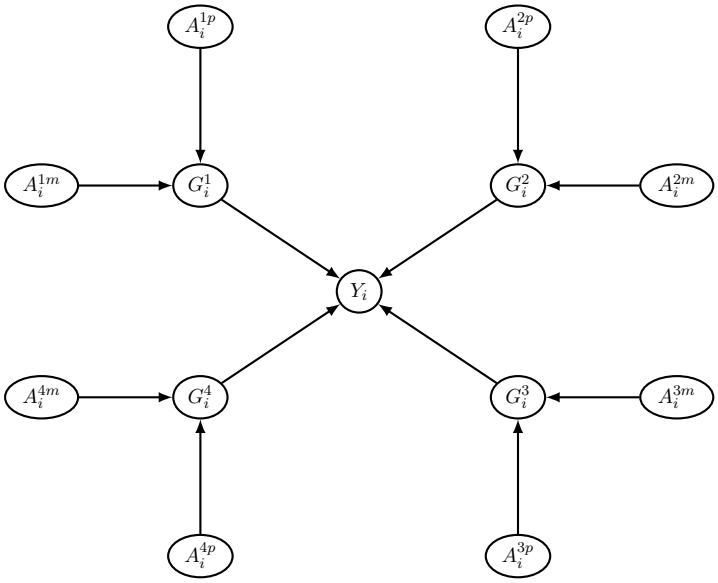


Figure 8.2: Allele DAG for a single individual with additional variables G_i^g, for $g = 1, \ldots, 4$, denoting the genotypes per gene.

between makers and a gene of interest is the core question in linkage analysis, assuming an independent segregation between MSH2 and MSH6 induces close to no bias on posterior carrier risk computations, in particular because of the low frequency of genotypes containing pathogenic mutations in both genes. Therefore most future computations will be done with the allele BN version for faster computations.

Equations (3.5) and (3.6) give the joint probability of the variables involved respectively in an allele and a selector BN. For convenience we include in both networks the set $X^g = \{X_i^g\}_{i=1,\ldots,n}$ of genotypes per gene $G^g \in \mathcal{G}$ such that graph parents of X_i^g are its corresponding paternal and maternal alleles $A_i^{g,p}$ and $A_i^{g,m}$ and, for all $i \in \{1, \ldots, n\}$, graph parents of Y_i become $\{X_i^g\}_{g\in\{1,\ldots,4\}}$. Figure 8.2 represents modifications involved in an allele DAG for a single individual. In our particular framework with four genes and state space reduction $\{X_i^g \in \{0, 1\}, X_i \in \mathcal{X}\}_{i=1,\ldots,n} \subseteq \text{ev}$ where ev is an evidence for the BN, adding individual genotypes involves same order of time complexity (and even tends to lower it over the majority of simulated families we tested) for exact inferences in most encountered pedigree structures. Note however that $X_i^1, \ldots, X_i^4$ all belong to the scope of common potential as they are graph parents of Y_i, hence adding an overall genotype $X_i = (X_i^g)_{g=1,\ldots,4}$ per individual with graph parents $\{X_i^g\}_{g=1,\ldots,4}$ is not advisable as this would lead to the unnecessary creation of n variables of cardinality $|\{0, 1, 2\}|^4$.

Moreover, we implemented the min-fill heuristic for determining the variable elimination orderings as it empirically leads to lowest computational complexities for most encountered families. As an example, the minimum, maximum and median of resulting complexities shown in Table 8.2 are computed from 20 runs of the algorithm using each heuristic over both families introduced in Section 3.2 and respectively pictured in Figure 3.7 and 3.8. The marriage loop in Family 3.8 explains the much

higher complexities obtained with the second family however marriage loops and consanguinity are exceptionally encountered.

Using same notation as in Section 3.2, the joint probability of variables involved in an allele BN with added genotypes per gene, under the assumption of independent segregation of genes, is given by the following adaptation of Equation (3.5) to our framework:

$$\mathbb{P}\big(\{A^{g,p},A^{g,m}\}_{g\in\{1,\dots,4},\{X^g\}_{g\in\{1,\dots,4\}},Y|S,\theta\big)= \prod_{i=1}^{n}\prod_{g=1}^{4}\mathbb{P}\big(A_i^{g,p}|A_{\mathrm{p}(i)}^{g,p},A_{\mathrm{p}(i)}^{g,m};q\big)\mathbb{P}\big(A_i^{g,m}|A_{\mathrm{m}(i)}^{g,p},A_{\mathrm{m}(i)}^{g,m};q\big)\mathbb{P}\big(X_i^g|A_i^{g,p},A_i^{g,m}\big) \times\mathbb{P}\big(Y_i|\{X_i^g\}_{g\in\{1,\dots,4\}},S_i;\alpha,\beta\big) \quad (8.4)$$

where, for $h\in\{p,m\}$, $A^{g,h}=\{A_i^{g,h}\}_{i=1,\dots n}$ such that, for all $i\in\{1,\dots,n\}$, for all $g\in\{1,\dots,4\}$, $A_i^{g,p}$ (respectively $A_i^{g,m}$) is the paternal (respectively maternal) allele carried by individual i for gene G^g and X_i^g denotes the genotypes carried by Individual i for the gene G^g. We assume that, for all $i\in\{1,\dots,n\}$,

$$\mathbb{P}\left(Y_i|\{X_i^g\}_{g\in\{1,\dots,4\}},S_i;\alpha,\beta\right)=\mathbb{P}\left(Z_i|\{X_i^g\}_{g\in\{1,\dots,4\}},S_i;\alpha\right) \times\mathbb{P}\left(B_i|\{X_i^g\}_{g\in\{1,\dots,4\}};\beta\right)^{\mathbb{1}_{\{i\in\mathcal{T}\}}}. \quad (8.5)$$

Finally, taking into account linkage disequilibrium between MSH2 and MSH6, the joint probability of variables involved in a selector BN (Equation 3.6) adapted to our framework is given by

$$\begin{aligned}\mathbb{P}\left(\{A^{g,p},A^{g,m}\}_{g\in\{1,\dots,4\}},\{\mathrm{SA}^{g,p},\mathrm{SA}^{g,m}\}_{g\in\{2,3\}},\{X^g\}_{g\in\{1,\dots,4\}},Y|S,\theta^s\right)=&\\ \prod_{i=1}^{n}\Bigg[\prod_{g=1}^{4}\left\{\mathbb{P}\left(A_i^{g,p}|A_{\mathrm{p}(i)}^{g,p},A_{\mathrm{p}(i)}^{g,m},\mathrm{SA}_i^{g,p}\right)\mathbb{P}\left(A_i^{g,m}|A_{\mathrm{m}(i)}^{g,p},A_{\mathrm{m}(i)}^{g,m},\mathrm{SA}_i^{g,m}\right)\right\}^{\mathbb{1}_{\{i\in\overline{\mathcal{F}}\}}}&\\ \times\left\{\mathbb{P}\left(A_i^{g,p}|q\right)\mathbb{P}(A_i^{g,m}|q)\right\}^{\mathbb{1}_{\{i\in\mathcal{F}\}}}\times\left\{\mathbb{P}\left(X_i^g|A_i^{g,p},A_i^{g,m}\right)\right\}\Bigg]&\\ \times\left\{\prod_{h\in\{p,m\}}\mathbb{P}\left(\mathrm{SA}_i^{2,h}\right)\mathbb{P}\left(\mathrm{SA}_i^{3,h}|\mathrm{SA}_i^{2,h};\delta\right)\right\}^{\mathbb{1}_{\{i\in\overline{\mathcal{F}}\}}}&\\ \times\mathbb{P}\left(Y_i|\{X_i^g\}_{g\in\{1,\dots,4\}},S_i;\alpha,\beta\right)\quad(8.6)\end{aligned}$$

where for all $i\in\overline{\mathcal{F}}$, $h\in\{p,m\}$ and $g\in\{2,3\}$, $\mathrm{SA}_i^{g,h}$ is the selector of allele $A_i^{g,h}$ and, for $g\in\{1,4\}$, $\mathrm{SA}_i^{g,h}=\emptyset$. The parameter δ is the genetic distance (Haldane, 1919), in Morgan units, between MSH2 and MSH6.

	min-fill heuristic				weighted min-fill		
	cliques	τ	complexity		cliques	τ	complexity
min	81	13	50,112	min	79	13	56,896
max	82	14	66,368	max	81	14	93,760
median	82	14	58,304	median	80	14	77,056
	min-weight heuristic				min-neighbors		
	cliques	τ	complexity		cliques	τ	complexity
min	78	14	64,128	min	79	14	64,576
max	82	20	1,405,344	max	82	17	217,120
median	80	15	117,664	median	80	15	99,776

(a) Family with no marriage loop nor consanguinity represented in Figure 3.7

	min-fill heuristic				weighted min-fill		
	cliques	τ	complexity		cliques	τ	complexity
min	76	20.0	4,502,784	min	74	20	4,321,024
max	76	20.0	4,502,784	max	76	26	85,118,720
median	76	20.0	4,502,784	median	75	24	21,101,824
	min-weight heuristic				min-neighbors		
	cliques	τ	complexity		cliques	τ	complexity
min	71	19.0	1,550,976	min	72	18	936,320
max	76	27.0	153,583,232	max	76	26	72,452,480
median	74	23.5	18,963,840	median	74	23	14,113,152

(b) Family with a marriage loop represented in Figure 3.8

Table 8.2: Minimum, maximum and median of the number of cliques (cliques), the treewidth (τ) and the associated time complexity (complexity) of computed junction-trees defined by a variable elimination over the allele BN with additional genotypes per gene associated with the family with no marriage loop nor consanguinity represented in Figure 3.7 (on the top, Figure 8.2a) and the family with a marriage loop represented in Figure 3.8 (at the bottom, Figure 8.2b). Results are obtained from 20 runs of the sum-product algorithm using the min-fill (top-left of each subfigure), the weighted min-fill (top-right), the min-weight (bottom-left) or the min-neighbors heuristic (bottom-right).

8.2.2 Conditional probability distributions and parameters

In this section we detail CPDs involved in the allele and selector BN. CPDs are defined before entering any evidence in the BN, in particular, before any state space reduction. Therefore, in this section, for all $i \in \{1, \ldots, n\}$ and $g \in \{1, \ldots, 4\}$, $X_i^g \in \{0, 1, 2\}$. LynchRisk's parameters are extracted from estimates available in recent literature. The survival parameter is adjusted with a model selection detailed in Section 8.3.4.

	$A_i^{g,p}=0$, $A_i^{g,m}=0$	$A_i^{g,p}=1$, $A_i^{g,m}=0$	$A_i^{g,p}=0$, $A_i^{g,m}=1$	$A_i^{g,p}=1$, $A_i^{g,m}=1$
$X_i^g=0$	1.0	0.0	0.0	0.0
$X_i^g=1$	0.0	1.0	1.0	0.0
$X_i^g=2$	0.0	0.0	0.0	1.0

Table 8.3: CPDs of the genotypic component of the form $\mathbb{P}(X_i^g|A_i^{g,p},A_i^{g,m})$.

8.2.2.1 Genotypic component

For given parameters $q=(q_1,q_2,q_3,q_4)$ and δ, most CPDs of the genotypic component have previously been detailed in Section 3.2.2 except those of the form $\mathbb{P}(X_i^g|A_i^{g,p},A_i^{g,m})$ given in Table 8.3 for a given $i\in\{1,\dots,n\}$ and $g\in\{1,\dots,4\}$. LynchRisk's parameter q is estimated from the proportion of heterozygous carriers in GP estimated by Win et al. (2017), i.e. $q=(0.051/2,0.035/2,0.132/2,0.140/2)$ and $\delta=5.24\times10^{-3}$cM (centiMorgan) is computed from the physical location (in base pair on chromosome 2) of MSH2 and MSH6 reported by the Genome Data Viewer[1] and the conversion from physical location to genetic distance (in Morgan) along chromosome 2 downloaded from the website of 1000 Genome Project[2].

8.2.2.2 Phenotypic clinical component

Each phenotype Y_i is composed of a time-to-event data Z_i associated with a CPD of the form $\mathbb{P}(Z_i|X_i,S_i;\alpha)$ and, for $i\in\mathcal{T}$, a clinical variable B_i associated with a CPD of the form $\mathbb{P}(B_i|X_i,\beta)$. In this section we develop CPDs of the form $\mathbb{P}(B_i|X_i,\beta)$ and in the next one, those of the form $\mathbb{P}(Z_i|X_i,S_i;\alpha)$.

We assume that clinical and biological tests are independent conditional on the genotype except MLH1 promoter hypermethylation and BRAF V600E mutation whose conditional independency is unsure. If both tests are performed on the same tumor, we solely consider the result for MLH1 promoter hypermethylation. For $i\in\mathcal{T}$, let $\{T_i^1,\dots,T_i^I\}\subseteq\mathcal{B}_i$ be the subset of tests carried out for individual i (except BRAF if a result on hypermethylation is available), we assume that

$$\mathbb{P}(B_i|X_i;\beta)=\prod_{j=1}^{I}\mathbb{P}(T_i^j|X_i;\beta).$$

Recalling that the sensitivity (respectively specificity) of a test is defined as the probability of a positive (respectively negative) test conditional on a genotype consistent with a positive (respectively negative) test, CPDs of the form $\mathbb{P}(T_i^j|X_i;\beta)$ are parametrized by β, the set of sensitivity and specificity per test. Note that one can reverse edges $X_i\to B_i$ in the DAG and conserves Markov property leading to CPDs of the form $\mathbb{P}(X_i|B_i;\tilde{\beta})$, parametrized by $\tilde{\beta}$, the predictive positive value and predictive negative value of biological tests. However sensitivity and specificity being more commonly used and estimated we set edges such that X_i is the parent of B_i.

[1] `https://www.ncbi.nlm.nih.gov/genome/gdv/browser`

[2] `https://www.internationalgenome.org/home`

LynchRisk's sensitivity and specificity of biological tests are estimates reported by Assasi et al. (2016) in CRC with the Pentaplex of five markers (respectively overall proteins) for MSI (respectively IHC). In the absence of reliable estimates for EC, we assume, in the first version of LynchRisk, that the sensitivity and specificity of MSI, IHC and MLH1 promoter hypermethylation in EC are equal those in CRC. That assumption is however not completely satisfying in particular because MSH6 is much more penetrant for EC than CRC and its association with an MSI status is weaker than those of MLH1 or MSH2. Estimates for EC will be updated when more data become available. Finally, in the absence of reliable estimates for the sensitivity and specificity of CRC location (proximal or distal), LynchRisk's estimates for that latter test are those of MMRpro, i.e. 0.873 and 0.625 respectively.

8.2.2.3 Phenotypic time-to-event component

Each variable Z_i is a time-to-event data describing the personal history of individual $i \in \{1, \dots, n\}$ regarding diagnoses in the set $\mathcal{D} = \{\text{CC}, \text{RC}, \text{EC}, \text{OC}, \text{GIC}, \text{UC}\}$. We assume that a given individual can encounter at most two diseases in his/her lifetime. That maximal number seems reasonable as it rules out partial information on exceptional phenotypes and avoid complex combinatorics. Therefore we consider the multi-state model represented in Figure 8.1 where State UN stands for "Unaffected" and for all $D \in \mathcal{D}$, State D stands for "Diagnosed with disease D" and state $\widetilde{D}$ sands for "Diagnosed with disease D after a disease in the set $\mathcal{D}$". Let $\mathcal{Z} = \{\text{UN}, \text{CC}, \text{RC}, \text{EC}, \text{OC}, \text{GIC}, \text{UC}, \widetilde{\text{CC}}, \widetilde{\text{RC}}, \widetilde{\text{EC}}, \widetilde{\text{OC}}, \widetilde{\text{GIC}}, \widetilde{\text{UC}}\}$ be the state space, the model is built under the following assumptions:

A 5. *Sex and genotype dependent transition intensities.*

For all $k, \ell \in \mathcal{Z}$, $s \in \{1, 2\}$ and $x \in \{0, 1, 2\}^4$, we denote by $\lambda_{k\ell}^{s,x}$ the transition intensity from state k to state ℓ conditional on sex s and genotype x.

A 6. *Piecewise constant transition intensities with common cuts $c = (c_1, \dots, c_M)$ where $c_1 = 25$, $c_M = 75$ and for all $j \in \{2, \dots, M = 11\}$, $c_j - c_{j-1} = 5$.*

Each transition intensity is assumed to be piecewise constant such that, for all $s \in \{1, 2\}$, $x \in \{0, 1, 2\}^4$ and $t \geq 0$,

$$\lambda_{k\ell}^{s,x}(t) = \mathbb{1}_{t \in]c_{j-1}, c_j]} \alpha_{k\ell,j}^{s,x} \tag{8.7}$$

where $c_0 = 0$, $c_{M+1} = c_{12} = \infty$ by convention and, for all $j \in \{1, \dots, M+1\}$, $\alpha_{k\ell,j}^{s,x}$ is the hazard rate between state $k \in \mathcal{Z}$ and state $\ell \in \mathcal{Z}$ in time interval $]c_{j-1}, c_j]$ for individuals of sex s and genotype x.

A 7. *Markov property.*

Under Markov property, transition intensities are assumed to be independent of the duration spent in a state. That assumption implies in particular that we ignore the age and duration dependent excess of risk of death after cancer per cancer type. Note however that, assuming that the excess of mortality rate due to cancer is additive on the hazard and independent of the genotype, considering death as a censoring

event induces no bias in posterior carrier probabilities. Nevertheless, LS carriers show better prognosis than non LS patients (Møller et al., 2017b). Moreover, ignoring the excess of risk of death does have an impact on future cancer risks computations and a semi-Markov model is a perspective for future versions of LynchRisk.

The parameter α is the set of vectors of hazard rates $\alpha_{k\ell}^{s,x} = (\alpha_{k\ell,j}^{s,x})_{j=1,\dots,M+1}$ for all $s \in \{1,2\}$, $x \in \{0,1,2\}^4$, $k,\ell \in \mathcal{Z}$ and is estimated from data reported in recent literature. That technical part is detailed in Section 8.3.4.

CPDs of the phenotypic time-to-event component are computed using the discretized Markov model introduced in Section 2.3. Let $\{Z(t), t \geq 0\}$ be a stochastic process where $Z(t)$ denotes the state occupied at time t, the time is discretized over $(i\Delta)_{i=1,\dots,N_i}$ steps of time with $\Delta = 1/12$ and N_i is such that $N_i\Delta$ is a choice of a maximal age (for instance 100 or age of last news for individual $i \in \{1,\dots,n\}$). Each variable Z_i is a Markov chain $Z_i = (Z_{i,1},\dots,Z_{i,N_i})$ such that for all $j \in \{1,\dots,N_i\}$, $Z_{i,j} = Z(j\Delta)$. Let $\text{ev}_{Z_i} = \cap_{j\in E\subseteq\{2,\dots,N_i\}}\{(Z_{i,j-1}, Z_{i,j}) \in \mathcal{Z}_{i,j-1}^* \times \mathcal{Z}_{i,j}^* \subset \mathcal{Z}^2\}$ be an evidence (or observation) for Z_i, CPDs of the form $\mathbb{P}(\text{ev}_{Z_i}|X_i = s, S_i = s; \alpha)$ are computed with the method detailed in Section 2.3 using Equation (2.5) in $\mathcal{O}(N_i \times |\mathcal{Z}|^2)$ time complexity. In practice, an age t reported on a pedigree indicates the entire year from birthday and computations are done at $t + \Delta$. In that sense, we adopt a conservative framework regarding posterior carrier risks computations as carriers of pathogenic mutation(s) tend to develop cancers at younger ages.

8.3 Transition intensities

The parameter α is the set of vectors of hazard rates $\alpha_{k\ell}^{s,x} = (\alpha_{k\ell,j}^{s,x})_{j=1,\dots,M}$ for $k,\ell \in \mathcal{Z}$, $s \in \{1,2\}$ and $x \in \{0,1,2\}^4$ and is estimated from available data in the literature. In the absence of individual data, several choices and additional assumptions had to be made. To the best of our knowledge, Weibull parameters in MMRpro are also estimated from exhaustive data reported in the literature which is listed but no details on the method applied for the parameter estimation are provided. In this section, we start with an overview of references used for estimating α, we then list and explain additional assumptions that must have been done and refine notation. Finally we develop the method used for estimating α using available data and expose other methods tried and ruled out.

8.3.1 Referenced data

Incidence data in carriers of a pathogenic mutation in MMR genes. Incidence data in carriers of MMR pathogenic variants are scarce and we decided to retain the work referenced by the InSiGHT group[3]. Dominguez-Valentin et al. (2020b) conducted an observational international multi-center prospective study and updated the Prospective Lynch Syndrome Database[4] composed of 6530 carriers of a single pathogenic MMR variant. The original database and inclusion criteria grouping centers from Finland, United Kingdom, Denmark, Spain, Germany, Norway,

[3] `https://www.insight-group.org`

[4] `http://lscarisk.org`

Sweden, Holland, Australia and Italy are detailed in (Møller et al., 2017a,b, 2018). Only carriers of class 4 or 5 variants according to the InSiGHT database[5] were included in (Dominguez-Valentin et al., 2020b). A deletion of the EPCAM gene, associated with epigenetic silencing of MSH2 (Ligtenberg et al., 2009), is assumed to be pathogenic MSH2. The authors report in supplementary material, the number of diagnoses per sex and localization and the observation years in five-years age cohorts from time interval $[25; 30[$ until time interval $[75; 80[$ in the following organs: colon, rectum, endometrium, ovary, stomach, small bowel, bile duct, gallbladder, pancreas, ureter/kidney, urinary bladder, prostate, breast and brain. We denote by $\mathcal{L}$ the set of cancers per organs considered by the authors. Estimates are also given per groups of organs colon/rectum, endometrium/ovary, upper gastrointestinal which aggregates stomach, small bowel, bile duct, gallbladder and pancreas, urinary tract which aggregates ureter/kidney and urinary bladder. The low penetrance of PMS2 for all LS-associated cancers leads to insufficient number of diagnoses and estimates reported for PMS2 carriers are pooled for both sexes.

Competing events are ignored, therefore for each organ (respectively group of organs), the first diagnosed cancer in that organ (respectively group of organs) counts, regardless the previous history of cancer in other localizations. Hence estimates are those of a two-state model as the one depicted in Figure 2.1 where State 1 stands for "Unaffected" and State 2 stands for diagnosed with the studied disease.

Incidence data in the general population (GP). Via a partnership between *Santé Publique France*, INCa, the biostatistics and bioinformatics department of civil hospices in Lyon and Francim, incidence and mortality rates per cancer are estimated from French registers every five years. In the latest version, volume 1, Defossez et al. (2019) estimate, among other quantities, annual incidence rates (number of new cases divided by person-time) per cancer type in 2018 as well as number of first diagnosed cancer per sex for 27 localizations including $\mathcal{L}$. The authors used registered data from 1990 to 2015 and a projection model detailed in the method section of their work. Competing risks are also ignored, therefore for each localization, new cases count in calculations regardless the past history for other cancer types. Estimates are reported per sex and localization, for $j \in \{1, \ldots, 19\}$, in time intervals $[a_{j-1}, a_j[$ such that $a_0 = 0, a_1 = 15$, $a_{19} = +\infty$ and for $j \in \{2, \ldots, 18\}$, $a_{j-1} - a_j = 5$.

8.3.2 Selected localizations

The localizations retained in LynchRisk are determined by available data in (Dominguez-Valentin et al., 2020b). Regarding estimated number of events and person years per organ and groups of organs considered by the authors, we decided to select cancers in colon (CC), rectum (RC), endometrium (EC) and ovary (OC) as well as groups of organs upper gastrointestinal (GIC) and urinary tract (UC) rather than separate component organs in order to keep sufficient data. Indeed, MMR genes being weakly penetrant in each component organs of these two latter groups, incidences are low

[5] `https://www.insight-group.org/variants/databases/`

and hard to estimate in separate organs. Furthermore we excluded brain du to insufficient data. As the inclusion of breast in the Lynch spectrum is controversial and confirmed by Dominguez-Valentin et al. (2020b), we decided to exclude breast. Finally as the relative risk of prostate cancer in carriers versus the GP is also low and because major fluctuation of annual incidences of prostate cancers in the French GP in recent years (Defossez et al., 2019), prostate was excluded. Consequently, as mentioned in previous sections, the set of Lynch-associated cancers retained in LynchRisk are $\mathcal{D} = \{\text{CC}, \text{RC}, \text{EC}, \text{OC}, \text{GIC}, \text{UC}\}$.

8.3.3 Additional assumptions

The first three additional assumptions listed thereafter are motivated by available data in the literature. In this section we explain limitations we faced and resulting assumptions that had to be made. Data reported in the literature are exhaustive and limited to aggregated number of events, person-years and/or annual incidence rates per time interval chosen by the authors with no access to individual data. Furthermore incidence data are reported as marginal quantities ignoring competing events between diseases. Consequently we make the following assumption:

A 8. *Diseases occur independently conditional on the genotype.*

Let T_i^k the time of first diagnosed cancer $k \in \mathcal{L}$ for an individual of sex $s \in \{1, 2\}$ and genotype $x \in \{0, 1, 2\}^4$, the hazard function of T_i^k in the absence of competing events is defined, for all $t \geq 0$ by

$$\lambda_k^{s,x}(t) = \lim_{\delta t \to 0} \frac{\mathbb{P}(t \leq T_i^k < t + \delta t | T_i^k \geq t, S_i = s, X_i = x)}{\delta t}$$

and we assume that

$$\lim_{\delta t \to 0} \frac{\mathbb{P}(t \leq T_i^k < t + \delta t | T_i^* \geq t, S_i = s, X_i = x)}{\delta t} \approx \lambda_k^{s,x}(t)$$

where T_i^* is the time of first diagnosed cancer in any localization $\ell \in \mathcal{L}$. Under such an assumption, we ignore in particular the dependency of diseases sharing common risk factors. That assumption seems however reasonable and is commonly made in genetic models such as BOADICEA (Antoniou et al., 2004) or MMRpro (Chen et al., 2006) prone to insufficient data. The assumption was partly verified with a two-by-two comparison of survival curves of the time to first diagnosed cancer computed from 0 to 75 years old in each group of organs considered by Dominguez et al. using either estimates of hazard rates for the group or the sum of estimates of hazard rates per component organs and assuming a piecewise constant hazard model with cuts chosen by the authors. In the absence of individual data, a rigorous comparison remains impossible but the maximal absolute error with respect to time computed per sex, mutated gene and group of organ ranges from 0.000 to 0.128 and from 0.000 to 0.054 when restricted to our chosen set $\mathcal{D}$. Table 8.4 offers a selection of computed maximal absolute errors at ages 40, 50, 60 and 75 per sex, mutated genes and group of organs considered by Dominguez-Valentin et al. (2020b). Errors are maximal for

	CRC				GIC				UC				Any			
	H1	H2	H6	S2	H1	H2	H6	S2	H1	H2	H6	S2	H1	H2	H6	S2
40	0.3	0.8	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	2.5	0.5	0.0
50	1.0	2.9	0.5	0.0	0.3	0.0	0.0	0.0	0.0	0.3	0.0	1.7	1.7	1.3	4.2	0.0
60	2.1	2.7	3.5	0.0	0.6	0.2	0.1	0.0	0.0	0.5	0.0	2.4	1.4	2.3	4.2	12.8
75	5.4	0.6	2.6	3.1	0.1	0.6	0.3	0.0	0.5	0.1	0.1	3.8	6.6	6.0	0.6	7.5

(a) Males

	CRC				EOC				GIC				UC				Any			
	H1	H2	H6	S2	H1	H2	H6	S2	H1	H2	H6	S2	H1	H2	H6	S2	H1	H2	H6	S2
40	0.1	0.3	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.8	1.1	4.0	0.0
50	0.3	2.2	0.4	0.0	0.0	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7	2.4	5.0	1.3	4.4
60	1.3	2.1	0.8	0.0	0.6	1.3	0.1	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	2.4	3.2	10.6	0.9	1.1
75	1.9	3.3	3.1	3.1	0.9	1.6	0.5	0.2	0.2	0.7	0.1	0.0	0.5	0.1	0.1	3.8	0.1	5.0	7.9	10.7

(b) Females

Table 8.4: Maximal absolute error (in percent) of the value taken by the survival function of the time to first diagnosis in groups of organs selected by Dominguez-Valentin et al. (2020b) computed at age 40, 50, 60 and 75 either with hazard rates associated with the group or the sum of hazard rates associated with component organs of the group. Computations are done for males (on the top) and for females (at the bottom) per gene denoted by its last letter and number for the sake of readability. CRC denotes colorectal cancer and aggregates CC and RC, EOC denotes gynecological cancer and aggregates EC and OC, GIC denotes upper gastrointestinal cancer and aggregates stomach, small bowel, bill duct, gall bladder and pancreas, UC denotes urinary tract cancer and aggregates ureter, kidney and urinary bladder and finally "Any" denotes the entire set $\mathcal{L}$.

carriers of a pathogenic variant in PMS2 which is much less penetrant than any of the other three genes rendering parameter estimation for PMS2 carriers difficult with wide confidence intervals du to insufficient data.

The previous assumption and the lack of estimates for incidences of subsequent cancers in the literature led us to make the following additional assumption:

A 9. *Incidences are not affected by past history of disease.*

We assume that the incidence rates a of subsequent cancer are equal those of first cancer per cancer type, sex and genotype. That assumption seems reasonable and confirmed in (Møller et al., 2017b) who reported higher but not significantly higher incidence of a subsequent cancer in $\mathcal{L}$ after a previous cancer in $\mathcal{L}$ for carriers of one pathogenic MMR variant in MLH1, MSH2 and MSH6. The authors could not report reliable results for carriers of pathogenic variants in PMS2 du to insufficient data.

As carriers of more than one pathogenic mutation are extremely rare, no incidence data are available in the literature for them. Consequently, we make the following additional assumptions:

A 10. *Dominant hazard model per gene. Additive effect of mutated genotypes on the hazard between genes.*

The dominant mode of inheritance of LS justifies the first part of the assumption. Under the assumption of an additive effect among carrier genotypes, we have, for all $s \in \{1,2\}$, $x = (x^1x^2x^3x^4) \in \{0,1,2\}^4 \setminus (0000)$ and $k \in \mathcal{L}$,

$$\lambda_k^{s,x} = \mathbb{1}_{\{x^1 \neq 0\}}\lambda_k^{s,(1000)} + \mathbb{1}_{\{x^2 \neq 0\}}\lambda_k^{s,(0100)} + \mathbb{1}_{\{x^3 \neq 0\}}\lambda_k^{s,(0010)} + \mathbb{1}_{\{x^4 \neq 0\}}\lambda_k^{s,(0001)}.$$

This assumption is also made by the authors of MMRPro. On the contrary the authors of BOADICEA select the major genotype in case of multiple pathogenic variants in different genes (Antoniou et al., 2002). Our assumption is motivated by the biological expression of each gene which is recessive and both alleles need to be inactivated for participating in the risk either by loss of heterozygosity or a pathogenic mutation in the functional allele. We assume that such inactivations occur independently leading to added instantaneous risks of mutation per remaining functional allele.

Note that under the above assumption, incidences conditional on a CMMRD genotype are assumed to be equal those conditional on a LS genotype. However, biallelic mutated genotypes being excluded from the set $\mathcal{X}$, entering the evidence $\{X_i \in \mathcal{X}\}$ in CPDs of the form $\mathbb{P}(Z_i|X_i, S_i; \alpha)$ sets values of the resulting potential to zero for all $X_i = x \notin \mathcal{X}$, whatever the initial value is.

Finally we add the following assumption:

A 11. *Null incidence for EC (respectively OC) after hysterectomy (respectively bilateral salpingo-oophorectomy).*

Hysterectomy and bilateral salpingo-oophorectomy is often proposed to patients of high risk of developing EC and/or OC and discussed for each patient according to psychological impact and parental project completion. The Manchester International Consensus Group recommend risk reduction surgery by hysterectomy and bilateral salpingo-oophorectomy for carriers of pathogenic variants in MLH1, MSH2 and MSH6 from 35-40 years old and outcomes of the surgery were evaluated by Dominguez-Valentin et al. (2020a). We assume that the incidence of EC (respectively OC) is null after hysterectomy (respectively bilateral salpingo-oophorectomy). That assumption seems reasonable regarding the rarity of ovarian remnant syndrome (Magtibay et al., 2005). Moreover Schmeler et al. (2006) studied risk reduction of EC and OC after prophylactic surgery in LS patients based on American registers of 315 LS women enrolled between 1975 and 2004. The authors reported no occurence of EC after hysterectomy and no occurence of OC after bilateral salpingo-oophorectomy in comparison with 33% of EC (respectively 5% of OC) for women of respective control groups (groups of women who did not receive surgery).

We assume that the effect of risk reduction surgery in cancelling incidences is independent of the genotype and therefore it can be ignored without introducing bias in posterior carrier risks computations. However it is taken into account for computing future cancer risks by setting appropriate transition intensities to zero from age at surgery. Note that hysterectomy alone also reduces the risk of OC

without cancelling it but estimates of hazard ratios are scarce in the literature and their inclusion is left for future versions of LynchRisk.

8.3.4 Estimation

Let us return to the expression of transition intensities in LynchRisk given in Equation (8.7). Under Assumption 9, for all $s \in \{1,2\}$, $x \in \{0,1,2\}^4$ and $k, \ell \in \mathcal{D} = \{\text{CC}, \text{RC}, \text{EC}, \text{OC}, \text{GIC}, \text{UC}\}$, we have $\lambda^{s,x}_{\text{UN}k} = \lambda^{s,x}_{\ell k'}$. In order to lighten notation, $\lambda^{s,x}_{\text{UN}k}$ will be denoted $\lambda^{s,x}_k$ from now on and, for $j \in \{1, \dots, M+1 = 12\}$, the hazard rate $\alpha^{s,x}_{\text{UN}k,j}$, from state H to state $k \in \mathcal{D}$ in time interval $]c_{j-1}, c_j]$ will be denoted $\alpha^{s,x}_{k,j}$. Under additional Assumption 10, LynchRisk's parameter α is reduced to the set of vectors of hazard rates $\alpha^{s,x}_k = (\alpha^{s,x}_{k,j})_{j=1,\dots,M+1}$ for $s \in \{1,2\}$ and $x \in \mathcal{X}^* = \{(0000), (1000), (0100), (0010), (0001)\}$.

In this section we detail the method used fo estimating α with data listed in Section 8.3.1. The method is applied per sex and localization, and therefore, in order to lighten notation, indexes and exponents s and k are removed from notation such that, for all $x \in \mathcal{X}^*$ and $j \in \{1, \dots, M+1\}$, $\lambda^{s,x}_k$ (respectively $\alpha^{s,x}_k$ and $\alpha^{s,x}_{k,j}$) is denoted λ^x (respectively α^x and α^x_j).

8.3.4.1 Hazard functions in the general population

The hazard function of the time to first diagnosis of disease $k \in \mathcal{D}$ for individuals of sex $s \in \{1,2\}$ in the GP is denoted λ and we assume it to be piecewise constant with cuts c such that

$$\lambda(t) = \mathbb{1}_{t \in]c_{j-1}, c_j]} \alpha_j$$

where α_j is the hazard rate of first disease k in GP of sex s in time interval $]c_{j-1}, c_j]$. For all $j \in \{2, \dots, M\}$, an estimate of α_j is assume to be estimated incidence reported by Defossez et al. (2019) for the corresponding organ and sex. An estimate of α_1 is given by the sum of the number of events divided by the sum of person-years in time intervals $[0, 15[$, $[15, 20[$, $[20, 25[$ reported by the authors. Moreover we assume that $\alpha_{M+1} = \alpha_{12}$ is given by incidence estimated by the authors in time interval $[75, 80[$. Finally, under Assumption 8, for each time interval $]c_{j-1}, c_j]$, the hazard rate associated with a group of organs is approximated by the sum of hazard rates of component organs of the group in the corresponding time interval.

8.3.4.2 Hazard rates in carriers of pathogenic MMR variants.

In genetic analyses, it is common to assume that hazard functions of the time to first diagnosis conditional respectively on a carrier genotype $x \in \mathcal{X}^* \setminus (0000)$ and a non-carrier genotype (0000), are linked, for all $t \geq 0$, by the equation $\lambda^x(t) = \text{RH}^x_0(t)\lambda^{(0000)}(t)$ where RH^x_0 is the time dependent hazard rate between carriers of genotype x and non-carriers. In the absence of incidence data for non-carriers, we consider a proportional hazard with age-dependent effects such that, for all $t \geq 0$ and $x \in \mathcal{X}^* \setminus (0000)$,

$$\lambda^x(t) = \text{RH}^x(t)\lambda(t)$$

where λ is the hazard function of the time of first diagnosis in the GP and RH^x is the time dependent hazard ratio between carriers of genotype x and the GP. We assume that RH^x is piecewise constant with cuts $d = (d_1, \dots, d_m) \subseteq c = (c_1, \dots, c_M)$ and we denote by $\rho^x = (\rho^x_1, \dots, \rho^x_{m+1}) \in \mathbb{R}^{m+1}_+$ the vector of hazard ratios such that for all $i \in \{1, \dots, m+1\}$, $\rho^x_i = \mathbb{1}_{t \in]d_{i-1}, d_i]} \mathrm{RH}^x(t)$ where $d_0 = 0$ and $d_{m+1} = +\infty$ by convention.

In the absence of incidence data in carriers of a pathogenic mutation in time interval $]c_0 = 0, c_1 = 25]$, we assume that $d_1 = c_1$ and $\rho^x_1 = 1$. Let $\theta = (\theta_2, \dots, \theta_{m+1}) = \big(\log(\rho^x_2), \dots, \log(\rho^x_{m+1})\big) \in \mathbb{R}^m$ be the unconstrained parameter of log hazard ratios, the log-likelihood of θ is given by (see Aalen et al., 2008):

$$\ell(\theta | c, d) = \sum_{i=2}^{|d|+1} \sum_{j, c_j \in]d_i, d_{i+1}]} \Big(A_j \big[\theta_i + \log(\alpha_j)\big] - B_j e^{\theta_i} \alpha_j \Big) \tag{8.8}$$

where, for $j \in \{2, \dots, M\}$ (respectively for $j = M+1$), A_j (respectively B_j) is the number of diagnosis (respectively person-years) in time interval $[c_{j-1}, c_j[$ (respectively $[75, 80[$) estimated by Dominguez-Valentin et al. (2020b) and α_j is the hazard rate in the GP detailed in Section 8.3.4.1. Note that we have a close formula for maximizing $\mathcal{L}(\theta)$ as

$$\arg\max_\theta \ell(\theta | c, d) = \left(\log \left[\frac{\sum_{j, c_j \in]d_i, d_{i+1}]} A_j}{\sum_{j, c_j \in]d_i, d_{i+1}]} B_j \alpha_j} \right] \right)_{i=2, \dots, |d|+1}. \tag{8.9}$$

Incidence data in carriers of a pathogenic MMR variant reported in (Dominguez-Valentin et al., 2020b) are prone to overfitting issue. We decided to perform a model selection in order to select breakpoints for θ. We chose a greedy descending stepwise selection associated with a likelihood ratio test with a p-value at 1% using Algorithm 5. Initial values for $d \subseteq c$ are such that d is of maximal size and ρ contains only finite values (i.e. we exclude null initial values for RH^x). Initial hazard ratios, i.e. computed from crude data in (Defossez et al., 2019; Dominguez-Valentin et al., 2020b) and estimates of hazard ratios after model selection are represented per sex, localization and mutated gene in Figure 8.3 for males and 8.4 for females.

Remarks about other methods. A BIC criteria was excluded for model selection as the number of observations is unclear in our framework and particularly in the absence of individual data. We thought of regularization methods such as a fused ridge regularization written as

$$\ell(\theta) = \sum_{j=1}^{M+1} \Big(A_j \big[\theta_j + \log(\alpha_j)\big] - B_j e^{\theta_i} \alpha_j \Big) - \sum_{j=2}^{M+1} \kappa \left(\theta_j - \theta_{j-1}\right)^2$$

where θ is of size $M+1$ and κ is a chosen penalty. However, such methods remain unsuitable for our context as a cross-validation, hence the choice of a penalty, is precluded in the absence of individual data.

Algorithm 5: Model selection with greedy descending algorithm

A: vector of number of events;
B: vector of person years of same size as A;
c: vector of cuts of size $|A| + 1$;
$d \subseteq c$: initial vector of cuts for θ;
x: i-th percentile of a chi square distribution with one degree of freedom (chosen i);
while TRUE **do**
  $\widehat{\theta^{(1)}} \leftarrow \arg\max_\theta \mathcal{L}(\theta|c, d)$ using Equation (8.9);
  $\ell^{(1)} \leftarrow \mathcal{L}(\widehat{\theta^{(1)}}|c, d)$ using Equation (8.8);
  $\ell^{(0)} \leftarrow (\ell_1^{(0)}, \ldots, \ell_{|d|}^{(0)})$ empty vector of size $|d|$;
  for j in $1, \ldots, |d|$ **do**
    $\ell_j^{(0)} = 2 * \left(\mathcal{L}(\arg\max_\theta \mathcal{L}(\theta|c, d_{\setminus j}) - \ell^{(0)}\right)$ s.t. $d_{\setminus j}$ is d offloaded of its j-th value
  end
  if $\min(\ell^{(0)}) > x$ **then**
    break
  end
  $d \leftarrow d_{\setminus \min_j\left(\ell_j^{(0)}\right)}$ update d by removing its i-th value where $i = \min_j(\ell_j^{(0)})$;
end
$\theta = \arg\max_\theta \mathcal{L}(\theta|c, d)$ using Equation (8.9);
return (d, θ)

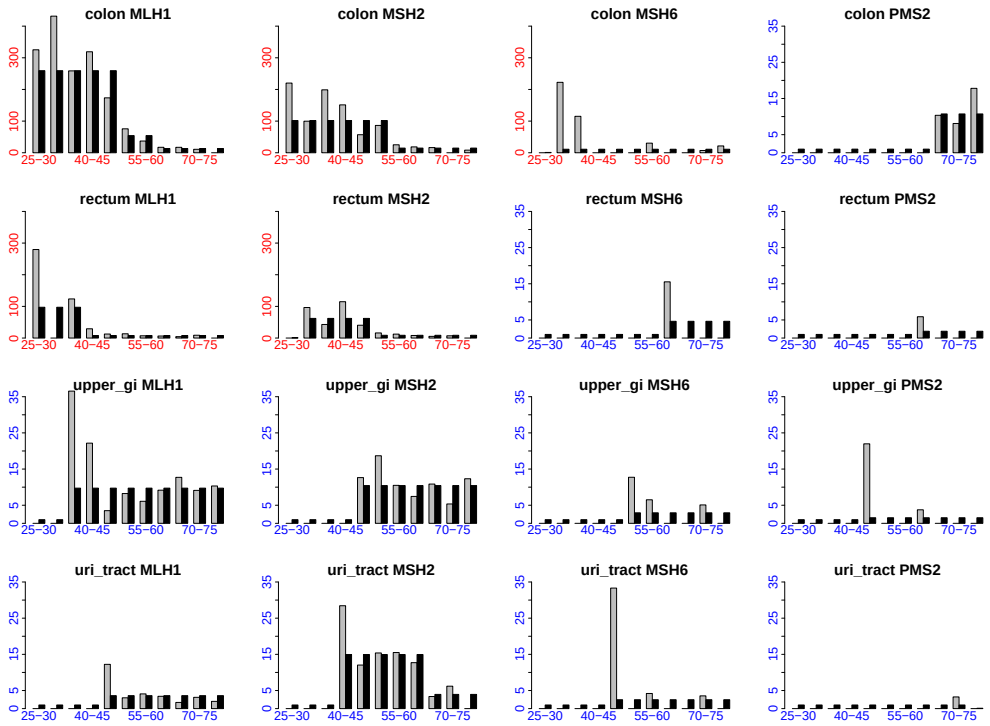


Figure 8.3: Initial hazard ratios computed from estimated incidence data in (Defossez et al., 2019; Dominguez-Valentin et al., 2020b) (grey) and after model selection in LynchRisk (black) for males. Scale ranges either from 0 to 400 (red axis) or from 0 to 35 (blue axis) according to genotype and localization.

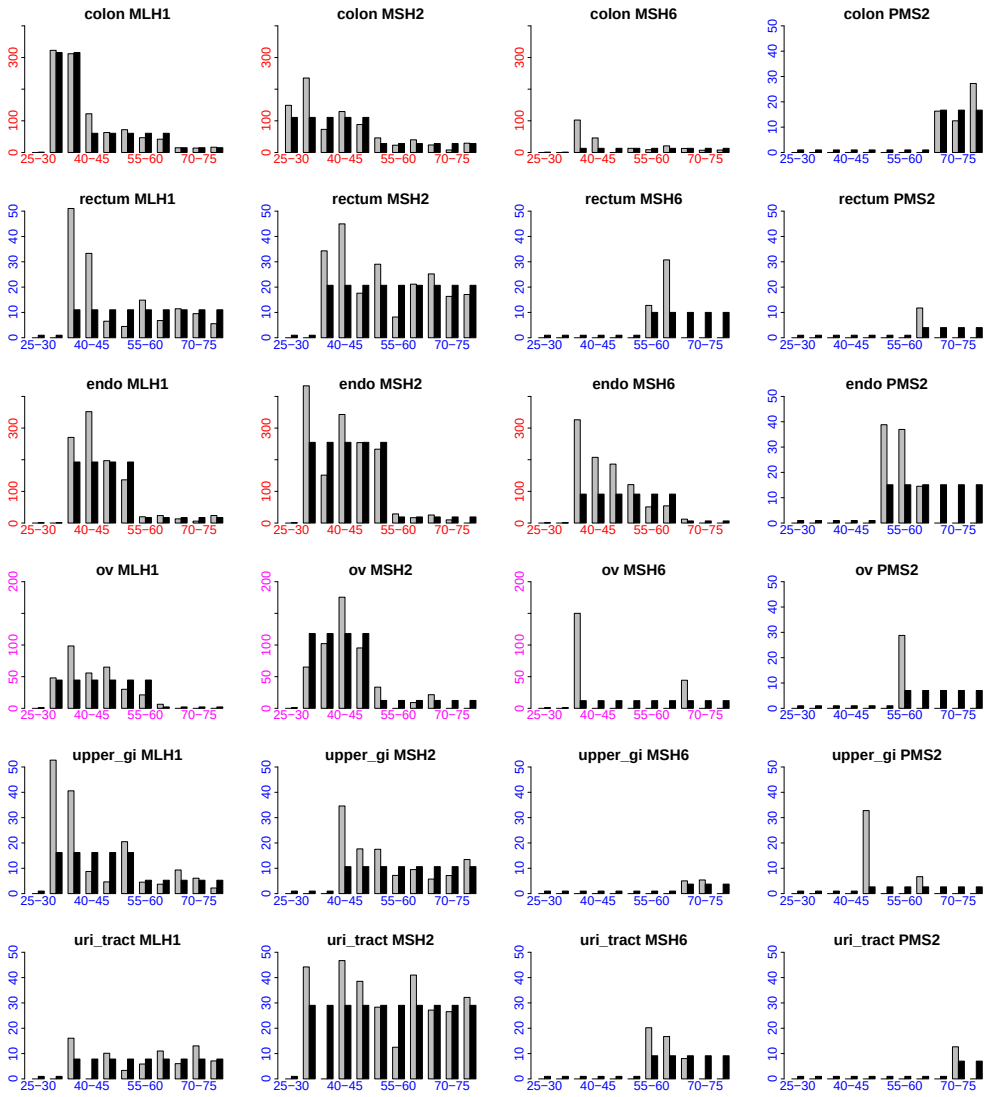


Figure 8.4: Initial hazard ratios computed from estimated incidence data in (Defossez et al., 2019; Dominguez-Valentin et al., 2020b) (grey) and after model selection in LynchRisk (black) for females. Scale ranges either from 0 to 400(red axis), from 0 to 200 (magenta axis) or from 0 to 50 (blue axis) according to genotype and localization.

8.3.4.3 Hazard rates in non-carriers

We assume that the state space of genotypes in the GP is $\mathcal{X}^* = \{(0000), (1000), (0100), (0010), (0001)\}$, i.e. the GP is solely composed of non-carriers and carriers of at most one pathogenic mutation in one MMR gene. That assumption seems reasonable in regard to the rarity of individuals carrying two or more germline pathogenic mutations. Let T_i be the time to first diagnosed cancer $k \in \mathcal{D}$ in the GP of sex $s \in \{1, 2\}$, the probability density function of T_i is a finite mixture model given, for all $t \geq 0$, by

$$f(t) = \sum_{x \in \mathcal{X}^*} p^x f^x(t) \tag{8.10}$$

where p^x is the frequency of genotype $x \in \mathcal{X}^*$ in the GP and $f^x(t)$ is the probability density function of T_i conditional on genotype $X_i = x$. Consequently the hazard function of T_i is given by (see McLachlan and McGiffin, 1994, for details):

$$\lambda(t) = \frac{1}{S(t)} \left(\sum_{x \in \mathcal{X}^*} p^x \lambda^x(t) S^x(t) \right)$$

where S is the survival function of T_i in the GP and λ^x (respectively S^x) is the hazard function of T_i (respectively the survival function of T_i) conditional on genotype $X_i = x$. Therefore assuming that λ and, for all $x \in \mathcal{X}^* \setminus (0000)$, λ^x are piecewise constant functions leads to a non piecewise function $\lambda^{(0000)}$ which is in contradiction with Assumption 6. However $\lambda^{(0000)}$ will be approximated by a PCH function with cuts $c = (c_1, \ldots, c_{M+1})$ and hazard rate in time interval $]c_{j-1}, c_j]$ denoted $\alpha_j^{(0000)}$. Recalling Equation 8.10, the survival function of T_i has the mixture form

$$S(t) = \sum_{x \in \mathcal{X}^*} p^x S^x(t). \tag{8.11}$$

For all $j \in \{1, \ldots, N+1\}$, $\alpha_j^{(0000)}$ is estimated by

$$\widehat{\alpha_j^{(0000)}} = \frac{1}{c_j - c_{j-1}} \log \left(\frac{S^{(0000)}(c_{j-1})}{S^{(0000)}(c_j)} \right)$$

as $\left[S^{(0000)}(c_{j-1})/S^{(0000)}(c_j)\right] = \exp\left[-(c_j - c_{j-1})\alpha_j^{(0000)}\right]$ in the framework of a PCH model. In practice, $S(c_{N+1})$ is assumed to be equal $S(80)$. Quantities of the form $S^{(0000)}(c_j)$ are computed with Equation (8.11) using genotype frequencies estimated by Win et al. (2017).

Cumulative distribution functions. We propose in Figure 8.5 a graphical representation of the cumulative distribution functions (CDFs) of the time first diagnosis per sex $s \in \{1, 2\}$, genotype $x \in \mathcal{X}^*$ and localization $k \in \mathcal{D}$ computed with estimates of hazard rates detailed in that section (plain lines) and initial values, i.e. crude hazard rates computed with number of events divided by person-years reported by Dominguez-Valentin et al. (2020b) (dashed lines). Computed CDFs with crude hazard rates are also available online on the Prospective Lynch Syndrome Database

website[6]. Genotypes are denoted per mutated gene for the sake of readability. The few remarks proposed thereafter are only qualitative and are intended to highlight the fact that results are consistent with expected values mentioned in the various literature on LS (Vasen et al., 2013; Goodfellow et al., 2015). The high penetrance of MLH1 and MSH2 in Lynch-associated cancers explains the fact that pathogenic variants in these genes are more frequently encountered in Lynch-associated tumors, in particular CRC and EC, despite their lower frequency in the GP. MSH6 is less penetrant in all cancer types except EC but its implication increases with age (Hendriks et al., 2004). The penetrance of MSH6 is equivalent to those of MLH1 and MSH2 in EC (Goodfellow et al., 2015) rendering LS women at particularly high risk of EC along with CRC. PMS2 is much less penetrant in all localizations (Senter et al., 2008; Ten Broeke et al., 2018). Separating CC and RC seems important although it is usually aggregated into CRC in current models as all MMR genes seem highly more penetrant for CC than RC at all age. In particular, the cumulative risk of CC (respectively RC) for MLH1 and MSH2 carriers ranges from 37% to 51% (respectively 10% to 14%) in males and 53% to 54% (respectively 11% to 20%) in females at age 80. Furthermore MLH1 seems more penetrant than MSH2 for CC whereas it is the contrary for RC. Let us extend that remark by noticing that MSH2 seems more penetrant than MLH1 in all extra-colonic tumors which is consistent with main conclusions in (Vasen et al., 2001) in particular for urinary tract.

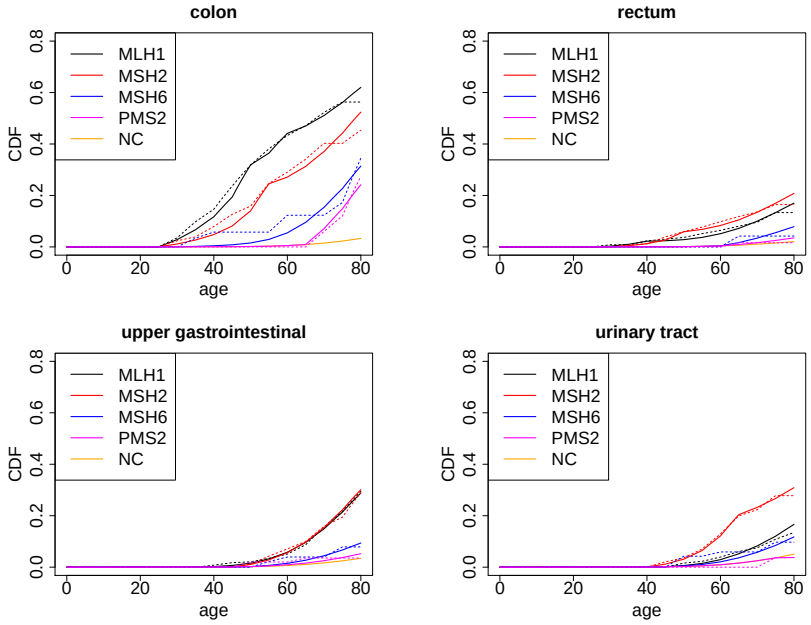
8.4 Computations

As a pedigree is a BN, one can apply the sum-product algorithm seen in Chapter 1 for computing exact posterior probabilities of interest in tractable time complexity for the high majority of encountered pedigrees (i.e. classical pedigrees and pedigrees with limited complex mating and consanguinity). In Section 1.1.1.2 we briefly adapt quantities seen in Chapter 1 to our context before presenting in Section 8.4.2 and 8.4.3 a set of chosen computed probabilities on simulated datasets in order to highlight the utility of the model. We finally discuss pros and cons of the model and perspective work in Section 8.5.

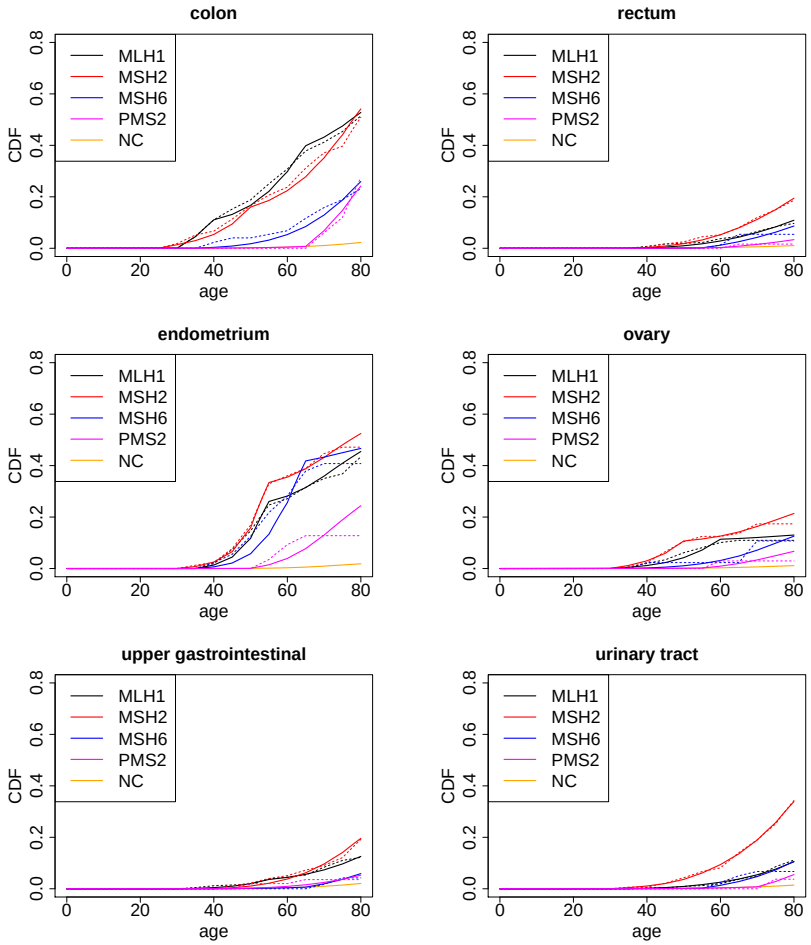
8.4.1 Evidence and potentials

In this section we briefly adapt quantities seen in Chapter 1 to our context with notation introduced in the previous sections. Let $\mathcal{B} = (G = (V = (V_1, \dots, V_p), \mathcal{E}), \mathbb{P})$ be a chosen (allele or selector) BN where V is the set of variables seen in Section 8.2.1 after pruning (see Section 1.6), $\mathbb{P}(V)$ factorizes over the product of CPDs $\mathbb{P}(V) = \prod_{u=1}^{p} \mathbb{P}(V_u|V_{\text{pa}(u)})$ where $\text{pa}(u)$ is the set of indexes of (graph) parents of V_u in G, $\text{pa}(\emptyset) = 0$ and $V_0 = \emptyset$. We denote by $\text{ev} = \{\text{ev}_{X_i}, \text{ev}_{Z_i}, \text{ev}_{B_i}\}_{i=1,\dots,n}$ an evidence for genotypes X_i and phenotypes $Y_i = \{Z_i, B_i\}$ such that, for all $i \in \{1, \dots, n\}$, $\text{ev}_{X_i} = \{X_i \in \mathcal{X}\}$, $\text{ev}_{Z_i} = \cap_{j \subseteq \{2,\dots,N_i\}} \{(Z_{i,j-1}, Z_{i,j}) \in \mathcal{Z}^*_{i,j-1} \times \mathcal{Z}^*_{i,j} \subset \mathcal{Z}^2\}$ if Z_i is not pruned and $\text{ev}_{Z_i} = \emptyset$ otherwise, $\text{ev}_{B_i} = \{B_i = b_i \in \{0,1\}^{|\mathcal{B}_i|}\}$ if $i \in \mathcal{T}$ and $\text{ev}_{B_i} = \emptyset$ otherwise.

[6] `https://sigven78.shinyapps.io/plsd_v4/`



(a) Male



(b) Female

Figure 8.5: Cumulative distribution functions of the time to first diagnosis in males (Figure 8.5a) and females (Figure 8.5b) per localization and genotype $x \in \mathcal{X}^*$ computed with estimated hazard ratios by model selection in LynchRisk (plain lines) and initial values (dashed lines). Legend: NC stands for non-carriers and carrier genotypes are denoted by the affected gene.

Individual posterior probabilities of carrying LS. The joint probability of V and ev is given by

$$\mathbb{P}(V, \text{ev}) = \prod_{u=1}^{p} \phi_u(S_{V_u})$$

where ϕ_u is the potential obtained by entering ev in $\mathbb{P}(V_u|V_{\text{pa}(u)})$ applying computational shortcuts if applicable (see Sections 1.1.1.2 and 1.6) and $S_{V_u} = \text{Scope}(\phi_u)$. Therefore one can apply the sum-product algorithm (see Chapter 1 from Section 1.3 until Section 1.7) to compute the posterior probability of any subset $V_U \subseteq V$, for instance, $\mathbb{P}(\{X_i^g\}_{g=1,\dots,4}|\text{ev})$ for the posterior probability that Individual $i \in \{1, \dots, n\}$ carries LS or $\mathbb{P}(X_i^g|\text{ev})$ for his posterior probability of carrying a deleterious variant in gene G^g.

In brief, let $\phi = \{\phi_u\}_{u=1,\dots,p}$ and σ be an elimination ordering over V (see Section 1.3.2 for choosing σ according to the factor graph H_ϕ), let $J = ((C_1, \dots, C_m), \mathcal{F})$ be a junction-tree defined by $\text{VE}(\phi, V, \sigma)$ (see Proposition 3), there exists $j \in \{1, \dots, m\}$ such that $\{X_i^g\}_{g=1,\dots,4} \subset C_j$. Indeed, each variable X_i^g, for $g \in \{1, \dots, 4\}$, is a graph parent of Z_i, hence there exist a potential containing $\{X_i^g\}_{g=1,\dots,4}$ in its scope. Applying Theorem 3 over clique C_j, the computational complexity to compute $\mathbb{P}(\{X_i^g\}_{g=1,\dots,4}|\text{ev})$ and $\mathbb{P}(X_i^g|\text{ev})$ for a chosen gene G^g is of the order of $\mathcal{O}\left(m \times \max_u |\mathcal{V}_u|^{\tau+1}\right)$ where $\mathcal{V}_u$ is the set of values taken by V_u and τ is the treewidth of J.

Distribution of the number of carriers of pathogenic variants and posterior probabilities conditional on that number. Applying the method introduced in Chapter 4, Section 4.2, we define the polynomial potentials, for all $i \in \{1, \dots, n\}$,

$$\boldsymbol{f}_{Z_i}(\{X_i^g\}_{g=1,\dots,4}) = \mathbb{P}(\text{ev}_{Z_i}|\{X_i^g\}_{g=1,\dots,4}) z^{\mathbb{1}\{(X_i^g)_{g=1,\dots,4} \in \mathcal{X} \setminus (0000)\}}$$

and for all $V_u \in V \setminus \{Z_i\}_{i=1,\dots,n}$, $\boldsymbol{f}_{V_u}(S_{V_u}) = \phi_u(S_{V_u})$. The unnormalized probability generating function of the number N of carriers is given by (see equation 4.3):

$$\text{pgf}_{N|\text{ev}}(z) = \sum_{k=0}^{\infty} \mathbb{P}(N = k, \text{ev}) z^k = \sum_{V} \prod_{u=1}^{p} \boldsymbol{f}_{V_u}(S_{V_u})$$

computed in $\mathcal{O}\big(n \times m \times \max_u |\mathcal{V}_u|^{\tau+1}\big)$ with an inward pass of the sum-product algorithm (see Equation 4.4). Similarly, for a chosen subset of variables $V_U \subset V$ one can compute $\sum_{k=0}^{\infty} \mathbb{P}(N = k, V_U, \text{ev}) z^k$ for the same time complexity with an additional outward pass.

Future cancer risks. Finally, future cancer risks for individual $i \in \{1, \dots, n\}$, i.e. future state space configuration for Z_i are computed with the method developed in Section 2.3 and in particular by a direct application of Equation 2.5 with the evidence $\text{ev} = \{\text{ev}_{X_i}, \text{ev}_{Z_i}, \text{ev}_{B_i}\}_{i=1,\dots,n}$ defined at the beginning of this section using two inward passes in $\mathcal{O}\left(m \times \max_u |\mathcal{V}_u|^{\tau+1}\right)$ each.

Another method was proposed in Nuel et al. (2017), an article published in *Computational and Mathematical Methods in Medicine* at the end of my Master's

thesis (see Appendix B), for obtaining the full posterior distribution of the time T of future disease in a competing risk setting between a disease and death. The probability of carrying a deleterious allele evolves with time while the individual remains free of disease and therefore, we express in this article the mixture form of the distribution of T. The extension to that method to a multi-state model with multiple diseases as the one implemented in LynchRisk is left for future work.

8.4.2 Contribution of personal histories of cancer to posterior probabilities

In order to explore the individual contribution of each phenotypic data to posterior risks computations via potentials of the form $\phi_{Z_i}(\{X_i^g\}_{g=1,\dots,4})$ and $\phi_{B_i}(\{X_i^g\}_{g=1,\dots,4})$, we propose in this section computed examples of posterior carrier probabilities in the absence of family history.

Computed posterior LS probabilities conditional on a first disease diagnosed at ages ranging from 35 to 75 and no other disease in the set $\mathcal{D}$ are listed in Table 8.5. Whenever it is appropriate probabilities conditional on a positive MSI result or a proximal or distal CC location are added. As expected, as all four genes are much less penetrant for GIC and UC, posterior carrier risks conditional on one such diagnosed cancer are low at all age. Note the highly increased posterior probability of carrying LS when conditioning on a positive MSI status and, in lower proportions, conditional on a proximal CC and a decreased risk when conditioning on a distal CC. The prevalence of LS conditional on a CC seems concordant with values reported by Vos et al. (2020) before 40 (18%) and overestimated for older individuals (1.7% between age 40 and 65 and 0.7% after 65). Note however that estimates of allele frequencies in LynchRisk are those reported by Win et al. (2017) with an overall LS prevalence at 0.358% in the GP. LS prevalence is however prone to controversies. For instance, Patel et al. (2020) report estimated ranging from 1/180 to 1/100 whereas Patel et al. (2020) estimate LS prevalence at 1/500. In MMRpro the estimated prevalence for LS is 0.23%.

LS women are of higher risk of EC than CC whereas it is the contrary in the general population leading to concordant computed posterior risks conditional on EC or CC. Indeed all three main genes (MLH1, MSH2 and MSH6) are highly penetrant for EC whereas pathogenic variants in MSH6 are much less involved in CC. Posterior carrier risks in EC women may be overestimated in LynchRisk when compared to expected values according to the literature as most articles report higher but close prevalence of LS in EC and CRC women at similar age. One of our first future goal is parameter, and in particular incidence rates, estimation. Conditioning on a MSI status leads, as expected, to increased posterior LS probabilities however we set the sensitivity and specificity of biological tests in EC as equal those in CRC which may be inappropriate. In particular, MSH6 being highly penetrant for EC and more weakly associated with MSI-H status than MLH1 or MSH2 are, we expect a significant difference of sensitivity and specificity of biological tests in EC and CRC. LynchRisk's parameters will be updated as soon as more data become available.

In Table 8.6 we report computed posterior carrier risks per gene conditional on

	Males					Females				
	35	45	55	65	75	35	45	55	65	75
CC	15.9	14.0	3.5	3.3	2.6	18.9	7.1	3.5	3.5	2.4
CC+MSI	43.8	40.2	12.9	12.2	9.9	48.9	23.8	13.1	12.9	9.3
CC+prox.	30.6	27.5	7.7	7.3	5.9	35.1	15.0	7.8	7.7	5.5
CC+dist.	15.9	14.0	3.5	3.3	2.6	18.9	7.1	3.5	3.5	2.4
RC	6.7	2.5	0.7	1.1	0.9	1.5	1.3	1.8	1.6	1.1
RC+MSI	23.0	9.5	2.8	4.4	3.4	5.8	5.0	7.1	6.1	4.3
EC	–	–	–	–	–	26.2	23.7	12.1	2.8	2.0
EC+MSI	–	–	–	–	–	59.4	56.1	36.1	10.6	7.6
OC	–	–	–	–	–	7.5	6.5	3.6	1.8	1.3
GIC	0.8	1.1	1.1	0.9	0.7	1.1	1.4	0.8	0.8	0.5
UC	0.4	1.1	0.9	0.6	0.3	1.6	1.4	1.8	1.2	1.3

Table 8.5: Posterior LS probabilities (overall genes) conditional on a diagnosed disease in the set $\mathcal{D}$ and additional clinical data and/or MSI DNA tumor status (in lines) at various at ages (in columns) for males (left columns) and females (right columns).

a diagnosed EC at ages ranging from 35 to 75 and no other diagnosis. Additional conditioning are added including results of IHC on MMR proteins and/or MLH1 promoter hypermethylation if appropriate. Results are consistent in particular with increased or decreased risks when conditioning on biological results. Surprisingly the posterior carrier risk in MLH1 is higher than MSH2 conditional on a diagnosis at any age whereas MSH2 seems more penetrant than MLH1 for EC (Figure 8.5b). This is explained by a higher frequency of MLH1 carriers than MSH2 carriers in the GP estimated by Win et al. (2017). Note that we could not test whether the differences in penetrance are statistically significant or not in the absence of individual data. Moreover Vasen et al. (2001) report higher risks of developing CRC and EC for MSH2 carriers when compared to MLH1 carriers but not significantly higher. Note the higher posterior probability of carrier a pathogenic variant in MSH6 at all age even if the cumulative distribution function of the time to first EC in MSH6 carriers is below the one in MLH1 and MSH2 carriers (Figure 8.5b) before age 62. Two facts can explain this. Firstly, pathogenic variants in MSH6 are more frequent in the GP than pathogenic variants in MLH1 or MSH2. Secondly MSH6 is much less penetrant for other Lynch-associated cancers rendering the probability of surviving free of any other cancer at any age less likely for MLH1 or MSH2 carriers than MSH6 carriers. The posterior probability of carrying a pathogenic variant in PMS2 in very low but increases with the age at diagnosis and even rises above those of the other three genes in elderly. This observation should however be tempered by the fact that PMS2 variants are more frequent in the GP. Finally a positive test on MLH1 promoter hypermethylation drops, as expected, posterior probabilities of carrying a pathogenic MLH1 variant, even below the frequency of heterozygous carriers in the GP after 55.

For the rest of the section, we compare results computed respectively with Lyn-

		$t = 35$	45	55	65	75
MLH1	EC_t	8.16	7.01	0.42	0.29	0.20
	EC_t + H1/S2 loss	33.63	29.70	2.14	1.35	0.93
	EC_t + H1/S2 loss + Hyper	2.41	2.02	0.11	0.07	0.05
MSH2	EC_t	7.39	6.16	0.28	0.20	0.09
	EC_t + H2/H6 loss	29.06	24.97	1.32	0.93	0.44
MSH6	EC_t	10.55	10.42	9.41	0.43	0.32
	EC_t +iso.H6 loss	39.68	38.71	33.51	2.02	1.49
PMS2	EC_t	0.16	0.16	1.97	1.89	1.36
	EC_t +iso.S2 loss	0.77	0.76	9.42	8.41	6.16

Table 8.6: Posterior probabilities (in percent) of carrying a deleterious variant per gene (left column) conditional on EC diagnosed at various ages t and/or results of IHC testing and/or hypermethylation of MLH1 promoter if appropriate. Legend: a loss of a protein or a dimer of proteins denotes the associated loss and no observed other loss. Hyper denotes an observed hypermethylation of MLH1 promoter in the tumor.

chRisk and MMRpro. $PREMM_5$ is not designed for risks computations in the absence of relevant family history and therefore a comparison with $PREMM_5$ in that section would make no sense but will be done in Section 8.4.3.

Computed posterior LS probabilities and posterior probabilities of carrying a deleterious variant in MLH1 or MSH2 for a male, conditional on personal histories of CC or RC at age 35, 45 or 55 are reported in Table 8.7. Various biological testing results are integrated in the calculation. Computations are done with LynchRisk (LR) and with MMRpro (Mp). Comparing the first and second column by sets of three columns, let us again lay emphasis on the importance of considering colon and rectal cancer separately as an individual diagnosed with CC is much more likely to carry LS than an individual diagnosed with RC at same age. Note the consistency of results with posterior carrier probabilities in all genes, in MLH1 and in MSH2 increased when conditioning on MSI status and increased risks of carrying a pathogenic variant in MSH2 (respectively MLH1) when conditioning on loss MSH2/MSH6 (respectively MLH1/PMS2 loss) and no other loss observed by IHC. Taking into account which protein shows lack of expression by IHC is of high importance and is ignored in MMRpro which aggregates any type of MMR deficiency (MSI status or loss of any protein observed by IHC) into a single valued observation. However one can see that considering each testing result appropriately leads to more precise conclusions when comparing results of the second, third and fourth sets of three rows. Additionally, conditioning on MLH1 promoter hypermethylation is another important data to integrate in calculation as suggested by all current guidelines with a highly decreased (respectively increased) risk of carrying a deleterious variant conditional on hypermethylation (respectively no hypermethylation) of MLH1 promoter. Alternatively, one can see that conditioning on a screened BRAF V600E mutation leads to similar posterior carrier risk than MLH1 promoter hypermethylation, both tests being equivalently valid.

	overall genes			MLH1			MSH2		
	C (LR)	R (LR)	CRC (Mp)	C (LR)	R (LR)	CRC (Mp)	C (LR)	R (LR)	CRC (Mp)
PH alone									
age 30	16.47	6.99	19.15	11.76	4.65	8.02	3.24	2.08	9.09
40	15.12	2.64	32.68	10.54	0.35	13.68	3.10	2.02	15.51
50	6.00	0.77	18.32	1.75	0.26	7.67	2.71	0.24	8.69
PH + MSI									
age 30	44.81	23.63	73.00	32.00	15.72	30.94	8.81	7.03	34.99
40	42.32	10.03	84.70	29.51	1.32	35.88	8.68	7.69	40.57
50	20.80	3.10	71.83	6.08	1.06	30.33	9.39	0.96	34.32
PH + MSH2/MSH6 loss and no other loss									
age 30	17.81	10.18	73.00	1.43	0.55	30.94	15.05	9.49	34.99
40	16.95	9.10	84.70	1.27	0.04	35.88	14.35	8.93	40.57
50	13.31	1.30	71.83	0.20	0.03	30.33	11.81	1.12	34.32
PH + MLH1/PMS2 loss and no other loss									
age 30	40.27	19.41	73.00	39.78	19.06	30.94	0.29	0.22	34.99
40	37.32	2.05	84.70	36.82	1.65	35.88	0.28	0.25	40.57
50	8.67	1.42	71.83	8.05	1.24	30.33	0.33	0.03	34.32
PH + MLH1/PMS2 loss, no other loss + MLH1 promoter hyper.									
age 30	3.92	1.57	NA	3.12	1.14	NA	0.46	0.27	NA
40	3.54	0.49	NA	2.76	0.08	NA	0.43	0.25	NA
50	1.10	0.25	NA	0.43	0.06	NA	0.35	0.03	NA
PH + MLH1/PMS2 loss, no other loss + No MLH1 promoter hyper.									
30	78.07	55.86	NA	77.89	55.67	NA	0.11	0.13	NA
40	75.85	8.57	NA	75.66	8.19	NA	0.11	0.24	NA
50	32.28	6.45	NA	31.82	6.28	NA	0.24	0.03	NA
PH + MLH1/PMS2 loss, no other loss + screened BRAF V600E									
age 30	3.07	1.25	NA	2.27	0.82	NA	0.46	0.27	NA
40	2.78	0.47	NA	2.00	0.06	NA	0.44	0.25	NA
50	0.98	0.23	NA	0.31	0.04	NA	0.35	0.03	NA

Table 8.7: Posterior carrier probabilities, in percent, for a male for overall genes (left columns) and the two most penetrant genes for the localization, MLH1 (middle columns) and MSH2 (right columns) conditional on a diagnosed cancer CC or RC (per column) at 30, 40 or 50 years of age (per line) computed with LynchRisk (LR) and MMrpro (Mp). Note that MMRpro aggregates colon and rectal cancer into colorectal cancer (CRC). PH means personal history of cancer, hence CC, RC or CRC at indicated age. Additional conditioning on biological testing results are integrated (per sets of three rows) where "hyper." stands for hypermethylation. Values are grayed if repeated, i.e. equal to quantities computed conditional on another evidence.

	CC_{35}	CC_{60}	UC_{40}	CC_{35}, CC_{60}	RC_{35}, CC_{60}	UC_{40}, CC_{60}
LR	15.91	2.12	0.81	**59.49**	**35.57**	**38.63**
Mp	24.38	1.99	NA	**24.38**	**24.38**	**1.99**

Table 8.8: Posterior probabilities of carrying LS (overall genes) in percent, conditional on various evidences of interest computed with LynchRisk (LR) and MMRpro (Mp). For $D \in \mathcal{D}$, D_t stands for diagnosed with disease D at age t. Posterior probabilities conditional on multiple diagnoses are bolded and repeated values (i.e. values previously computed when conditioning on another evidence) are grayed.

Finally in Table 8.8 we propose a selection of computed posterior probabilities of carrying LS in order to highlight the importance of taking into account multiple localizations and multiple events per individual, i.e. the importance of modeling time-to-event data in a multi-state model allowing for transitions between diseased states rather than a competing risk model as the one represented in Figure 2.2 which solely allows for taking the first diagnosis into account. In order to lighten notation, we use the common simplified notation in pedigree analysis (see Section 3.2.1) where D_t stands for diagnosed with disease $D \in \mathcal{D}$ at age t and we assume that the entire personal history of cancer for the studied diseases is reported such that, for instance, for $D^1, D^2 \in \mathcal{D}$, $PH = \{D^1_s, D^2_t\}$ stands for diagnosed with D^1 at age s and D^2 at age t and no other disease in the set $\mathcal{D}$. Comparing the first three with the last three columns, we notice as expected, that ignoring multiple diagnoses leads to severely underestimated posterior probabilities of carrying LS. We think that this is particularly important when studying LS as LS patients are not rarely victims of multiple cancers. Note also the importance of widening the spectrum of cancer types by considering the urinary tract cancer in the last column associated with phenotype $PH = \{UC_{40}, CC_{60}\}$.

8.4.3 Posterior risks conditional on a family history of disease

In this section we propose an overview of posterior risks computations conditional on a family history of cancer and/or biological testing results. Family history plus biological testing results are called an evidence. For readability, we will use the term family history denoted FH to design the set of available data, including biological testing results. Most quantities are computed with LynchRisk (LR), MMRpro (Mp) and $PREMM_5$ and compared. In this section, for the fluidness of the reading, personal histories of cancers are expressed with the conventional simplified notation in pedigree analyses introduced in Section 3.2.1. For instance $\{PH_i = UN_t\} \equiv \{Z_i(t) = UN, t \geq 0\}$ stands for an individual free of disease from birth up to age t and for $D \in \mathcal{D}$, $\{PH_i = D_t\} \equiv \{Z_i(s) = UN, s < t, Z_i(t) = D\}$ for an individual free of disease up to age t and diagnosed with disease D at age t. All family histories presented below are fictional.

Figures 8.6 and 8.7 are graphical representations of posterior carrier risks for each family member (on the right) conditional on various evidences (on the left). Quantities are computed with LynchRisk (black), MMRpro (red) and $PREMM_5$ (gray).

For readability, the scale vary from 0 to 70% for each graph except when conditioning on FH1 (all individuals are unaffected, scale 0-1%) and FH2 (one case, scale 0-10%). The posterior distribution of the number of carriers conditional on each family history is depicted in Figure 8.8. PREMM_5 being not designed for computing risks in the absence of cases in the family, plots associated with PREMM_5 conditional on FH1 are left empty. Note the magnitude of posterior probabilities when comparing LynchRisk/MMRpro and PREMM_5 and the importance of examining results qualitatively and determining a threshold. PREMM_5 returns low values conditional on any evidence (except in case of a single diseased individual as in FH2), so is the threshold for advocating genetic counseling consultation set appropriately at 5% by the National Comprehensive Cancer Network[7] (Boland et al., 2018) and even lowered at 2.5% by the authors.

Posterior LS probabilities computed with LynchRisk and MMRpro qualitatively agree when restricted to data handled by MMRpro (FH1 to FH4 and FH7).

Individual posterior probabilities conditional on FH1 are comparable between LynchRisk and MMRpro with coherent quantities regarding sex and age at last follow-up (the latter an individual survives free of disease, the lower his/her probability of carrying a deleterious variant). Moreover, females are at higher risk of disease as endometrial and ovarian cancers belong to the Lynch spectrum. Consequently, the latter a woman survives free of Lynch-associated cancer, the lower her probability of carrying a deleterious MMR variant when compared to a male of similar age. Greater values computed with LynchRisk could be partly explained by higher values of the vector of allelic frequencies in the general population (parameter q) as well as lower incidences in LynchRisk for colon and rectal cancer. However, this is in contradiction with more diseases considered in LynchRisk rendering individuals surviving free of considered diseases less likely to carry LS. Note however that LynchRisk's parameters are extracted and calibrated from literature published after 2016 and MMRpro has not been updated since 2006 which could explain discrepant results.

Conditioning on FH2 increases the posterior carrier probability for each family member, in particular Individual 9. Note the steeper increase of the risk of Individual 9 computed by PREMM_5 when compared risks computed by PREMM_5 conditional on other FH. This could be a consequence of PREMM_5 ignoring the detailed family history, in particular the detailed history of non-affected individuals and detailed relations between individuals. Therefore, the quantity computed for Individual 9 by PREMM_5 would have been equal in another family of individuals free of cancer, whatever the size of the family or its structure. Greater risks computed by MMRpro when compared to LynchRisk could be again a consequence of greater incidences for colorectal cancer in MMRpro as well as the lower number of considered disease. Hence, information such that Individual 9 and other family members also survived ovary, upper gastrointestinal and urinary tract cancer is lost in MMRpro. Results are consistance with increased risks in particular for Individual 6 and 2 despite an unaffected phenotype as, if a deleterious variant is transmitted to Individual 9, it must be carried by her ancestors. The models tend to favor a paternal transmission

[7] `https://www.nccn.org`

(from Individual 6) which could be explained by the overall consideration of unaffected phenotypes per family sides, in particular, ages of last news that tend to be lower in the left (on the picture) side of the family.

Conditioning on phenotype CC_{54} for Individual 2 reinforces the high posterior probability of a deleterious variant carried by Individual 2, 6 and 9, hence, similar posterior carrier probabilities for these three individuals even if Individual 6 is free of Lynch-associated cancer. Note the importance of considering the detailed family structure when comparing risks computed with PREMM_5 versus LynchRisk or MMRpro in FH2 versus FH3 with a steeper increase for Individual 6 (in the diagonal 2-6-9) and 5 (son of Individual 2) when computed with a pedigree-based model (LynchRisk or MMRpro) versus a logistic-regression based model (PREMM_5).

Adding a positive MSI tumoral status for Individual 9 is missed by PREMM_5 (equal risks computed in FH4 versus FH3) which is not designed for integrating biological testing results. On the contrary, such data increase risks computed by LynchRisk and MMRpro by about two fold. Carrier risk for Individual 5 is slightly below half the risk of Individual 2. Indeed, any child of Individual 2, have about half her risk of having inherited a deleterious allele (roughly, if we ignore carriers of two mutations) in the absence of phenotypic information. The unaffected phenotype at 54 for Individual 5 slightly lowers that relative risk as a Lynch-associated cancer by that age is more likely to happen in LS carriers.

MMRpro is not designed for considering subsequent cancers and therefore, posterior probabilities computed by MMRpro conditional on FH5 and FH4 are equal. LynchRisk and PREMM_5 on the contrary allow for adding subsequent cancers (up to one in LynchRisk) and risks computed conditional on FH5 for affected individuals are all above those computed conditional on FH3. Risk augmentation computed by LynchRisk is about three fold in carriers (compared to two folds when adding an MSI tumoral phenotype in FH4) suggesting that a survival model allowing for transitions between diseases is at least as important in risks computations than biological testing results (although such phenotypes are rarer in practice for the proband, i.e. the individual who seeks medical attention).

FH6 is equivalent to FH3 except for the phenotype of Individual 5 (GIC_54 instead of UN_{54}). A comparison between posterior probabilities computed conditional on FH6 and FH3 shows both the importance of including extra-colonic/extra-endometrial cancer in particular when studying the Lynch syndrome whose spectrum is large and the importance of modeling the structure dependency between genotypes in the family. Firstly, as MMRpro considers solely CRC and EC, the phenotype of Individual 5 in FH6 is equivalent to UN_{54} in MMRpro calculation, hence equal risks computed conditional on FH3 or FH6. On the contrary, phenotype GIC_{54} for Individual 5 leads to increased posterior probabilities of carrying LS for all family members on the left side of the family when computed with LynchRisk or PREMM_5. Secondly, as PREMM_5 ignores the precise family structure, relative risks between Individual 5 and 6 are overestimated by PREMM_5 whereas LynchRisk captures the higher posterior probability of carrying LS for Individual 6 when compared to Individual 5 as Individual 6 in on the diagonal 2-6-9.

FH7 is equivalent to FH3 with phenotype EC_{62} for Individual 7 and consequently

an increased risk for Individual 3, 4 and 8 and has little impact on posterior risks for 2, 5 and 6 as carriers of two mutations in different genes are allowed rendering two sides of the family possibly carriers. The use of PREMM_5 in such situation is ambiguous as only the affected side of the family should be considered. Results for PREMM_5 are here computed as if the whole dataset was considered. Note the higher risk of Individual 3 compared to Individual 2 despite a less protective phenotype UN_{77} versus UN_{78}. This is a consequence of EC and OC being included in the Lynch spectrum and in particular EC for which three genes are highly penetrant. Hence, a woman free of disease is less likely to be non-carrier than a man free of disease at all age and in particular in the elderly.

We see again the importance of considering several cancer types in particular when studying a syndrome of a large spectrum and the importance modeling precise dependence structure between all genotypes when comparing posterior risks conditional on FH8 versus FH7. Phenotype GIC_{54} for Individual 5 leads to increased risks in blood relatives of Individual 5 (left side of the family) computed with LynchRisk and PREMM_5. These increases are missed by MMRpro. Furthermore relative risks between Individual 5 and 6 is not captured appropriately by PREMM_5 who ignores the precise family structure whereas the higher probability of carrying LS for Individual 6 when compared to 5 is captured by LynchRisk. For the same reason, posterior risks for family members on the right side computed by PREMM_5 stay still when compared to FH7 whereas they are lowered by LynchRisk. Indeed, increased probabilities in the left side lead to decreased probabilities in the maternal side of Individual 9 as Individual 9 is more likely to carry at most one pathogenic mutation.

The distribution of the number of carriers denoted N conditional on each aforementioned FH is represented in Figure 8.8. As pathogenic variants are rare and cancer is a multifactorial disease with a high proportion of sporadic types, N takes value 0 with the highest posterior probability conditional on each FH except FH5 and FH8 containing respectively a case of multiple cancer for Individual 2 (FH5) and four diseased on eight individuals (FH8). Moreover, conditioning on zero or one affected individual (FH0 and FH1) leads to nearly null posterior probabilities of non-null values for N. Value one or two is very unlikely conditional on all FH as phenotype CC_{41} for Individual 9 from FH2 to FH8 renders her to be one of the most likely carrier of a pathogenic variant among family members, variant that must be carried at least by a parent and a grandparent. As expected the second most probable value for N is 3 when conditioning on FH3, 4 or 5 with Individuals 2, 6 and 9 and FH7 with Individuals 2, 6 and 9 or 3, 7 and 9 in the trio grandparent, parent, child. Third most probable value 4 is explained by adding the brother of 6 or 7 respectively. Conditioning on FH6 or 8 leads to second most probable value at 4 explained by high probability of carrying LS by Individuals 2, 5, 6 and 9.

Conditioning on the number of LS carriers allows one for targeting families at risk and highlighting individuals at risk as represented in Figure 8.9. In particular for instance in that example, the posterior probability of carrying LS for Individual 5 is highlighted when conditioning on $N = 4$ and $N = 2$ carriers. Note again in that example the higher posterior probability of carrying LS for Individual 3 when

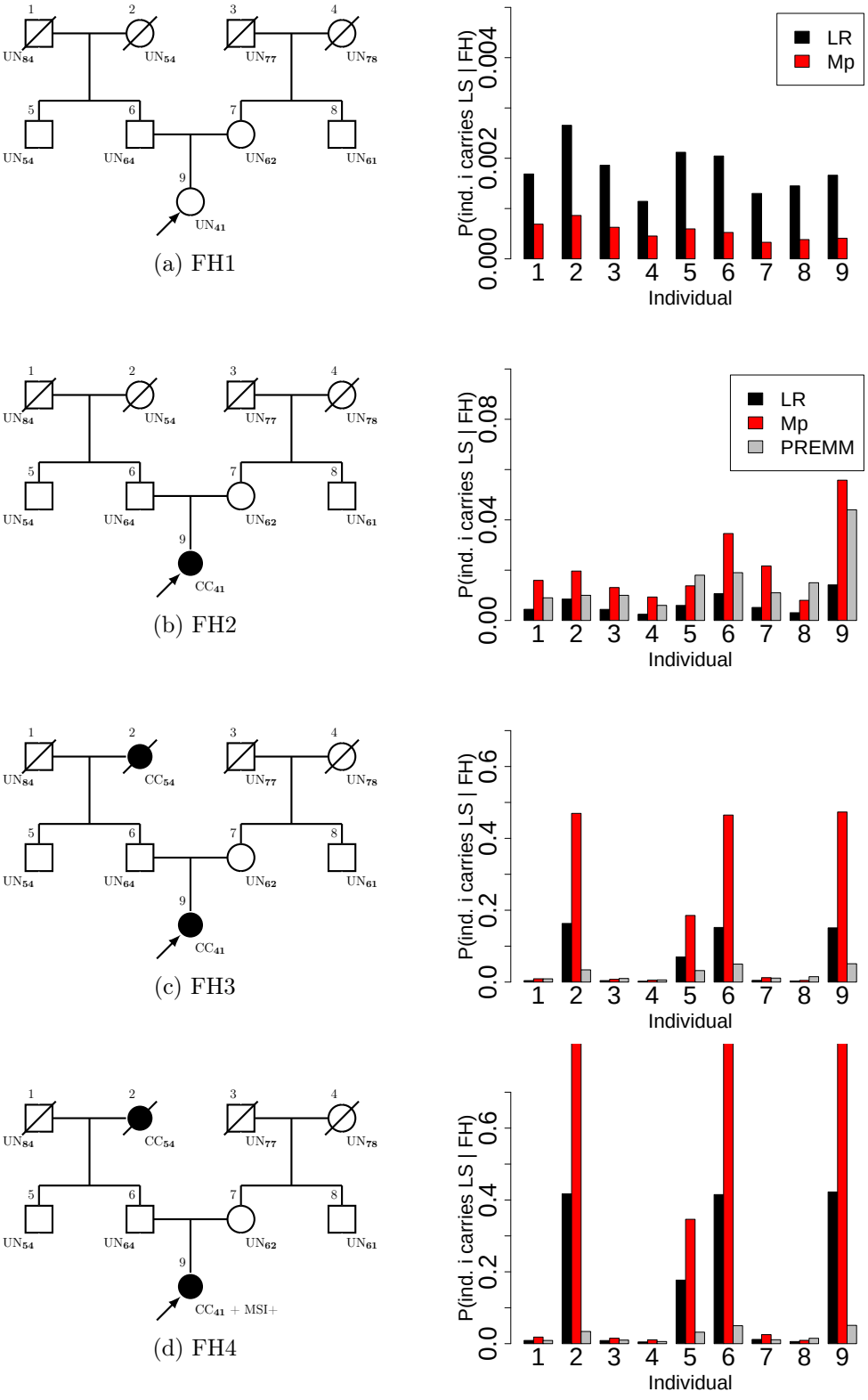
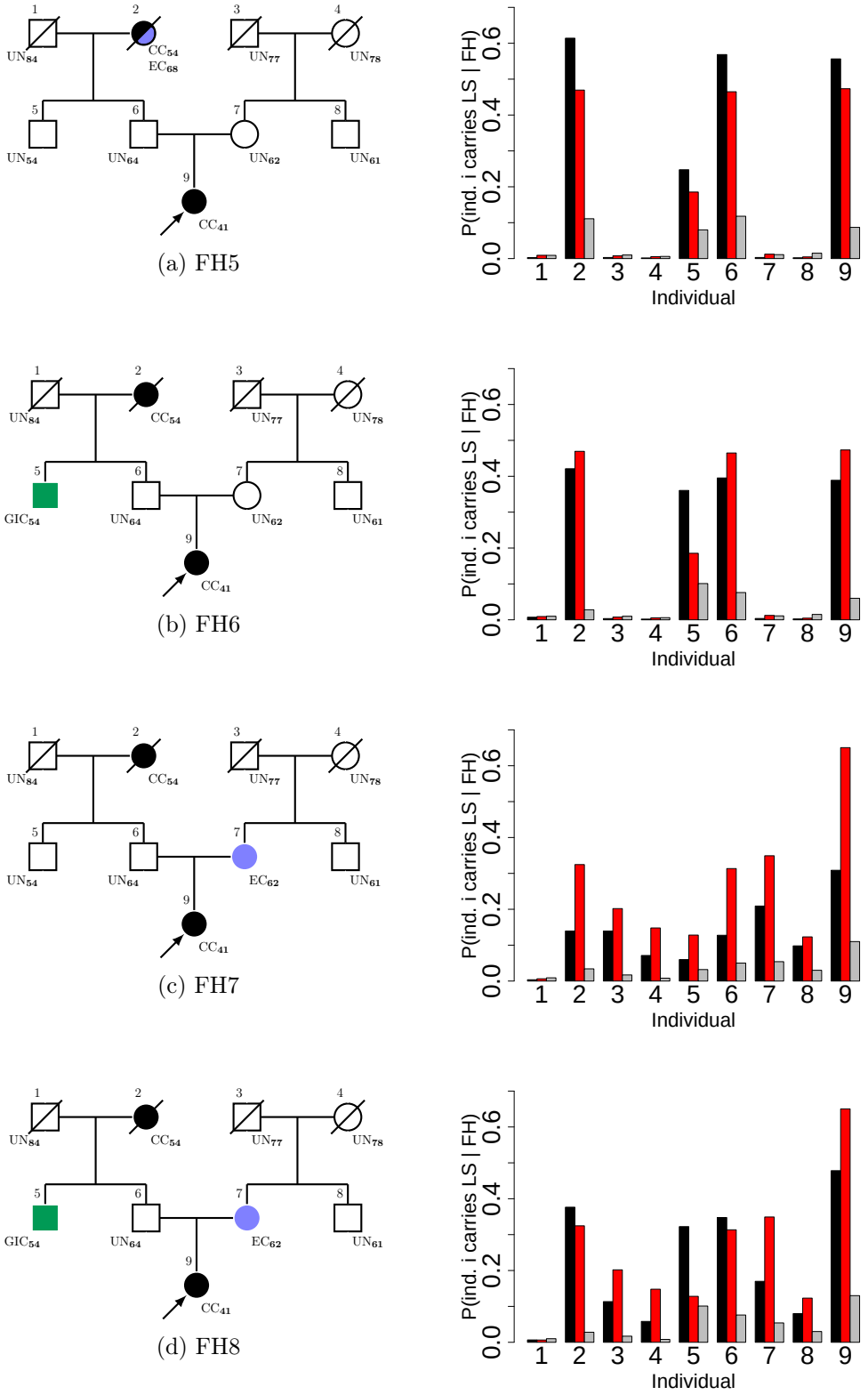


Figure 8.6: Individual posterior LS probability conditional on various FHs computed with LR (black), MMRpro (red) and PREMM$_5$ (gray). Scale vary from 0 to 0.005 (FH1), from 0 to 0.1 (FH2) and from 0 to 0.7 (FH3 and FH4). A filled black shape stands for an individual diagnosed with CC.



(a) FH5

(b) FH6

(c) FH7

(d) FH8

Figure 8.7: Individual posterior LS probability conditional on various FH computed with LR (black), MMRpro (red) and PREMM$_5$ (gray). Scale varies from 0 to 70. A filled black (respectively blue and red) shape stands for an individual diagnosed with CC (respectively EC and GIC).

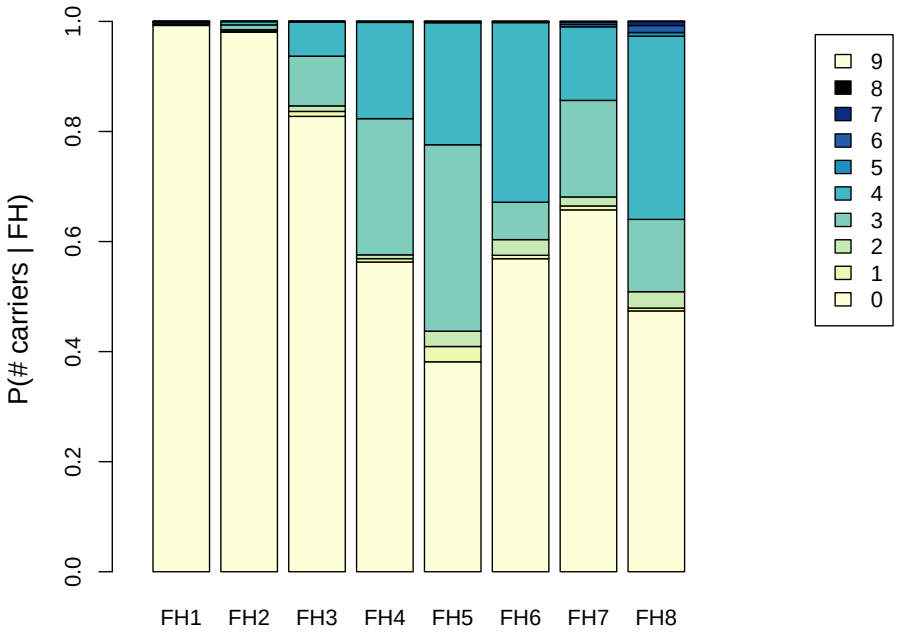
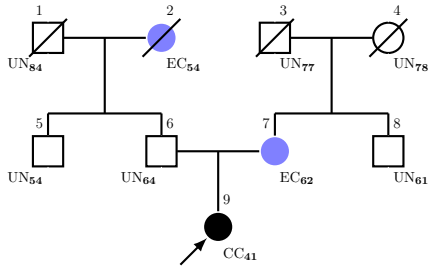


Figure 8.8: Distribution of the number of carriers conditional on family histories FH1 to FH8 in Figures 8.6 and 8.7.



		$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$
N	$\mathbb{P}(N\|\text{FH})$	$\mathbb{P}(X_i \neq (0000)\|N,\text{FH})$								
0	37.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	30.8	0.2	**80.5**	12.9	6.6	2.4	**80.6**	19.4	1.5	**96.0**
4	23.0	0.4	**79.1**	15.0	7.6	**77.4**	**77.6**	22.5	20.8	**99.7**
2	3.3	0.7	**83.7**	10.1	5.5	**58.9**	25.4	14.7	0.9	0.0
$\mathbb{P}(X_i \neq (0000)\|\text{FH})$		0.3	51.5	10.0	5.2	22.2	46.1	15.0	7.1	55.9

Figure 8.9: Posterior probability (in percent) of the four most probable number of carriers (left columns) and posterior risks of LS per individual conditional on that number and on the FH drawn on the top of the picture. The last line of the table gives the posterior LS probability per family member when conditioning solely on FH. Bolded values are the N highest ones.

compare to Individual 4 as women being at higher risk of LS are less likely to survive free of disease up to a great age.

Most previous examples are selected among severe family histories and show the importance of considering the exhaustive FH for posterior risks computations. We propose in Figure 8.10 a case report to account for the importance of considering unaffected individuals in risks computations. Individual 5 belongs to a large free of Lynch-associated cancer family and is himself diagnosed with a CC at age 54. His probability of carrying LS conditional on his phenotype regardless his FH is 5.49% and drops to 0.85% when conditioning on his FH. Adding respectively an MSI status or MSH2/MSH6 loss and no other MMR protein loss detected by IHC rises his posterior probability to 3.4% and 1.2% respectively. Similarly conditioning on MLH1/PMS2 loss, no other loss and no MLH1 promoter hypermethylation leads to a posterior LS probability at 3.3%. Hence even when conditioning on biological testing results positively associated with LS, his posterior LS probability is still below the one computed in the absence of FH.

Another example for highlighting the importance of the whole family history, including unaffected members is proposed with FH9 drawn in Figure 8.11 accompanied with computed posterior probabilities for each family member. We denote by $\text{FH9}_{\setminus i}$ the set of phenotypes in the family except the phenotype of individual i. In the hypothetical situation where Individual 1 were dead at age 28 instead of 79 (second line versus third line), we loose the information such that he survived free of disease from age 28 to age 79. Hence, he becomes more likely to be the person who transmit-

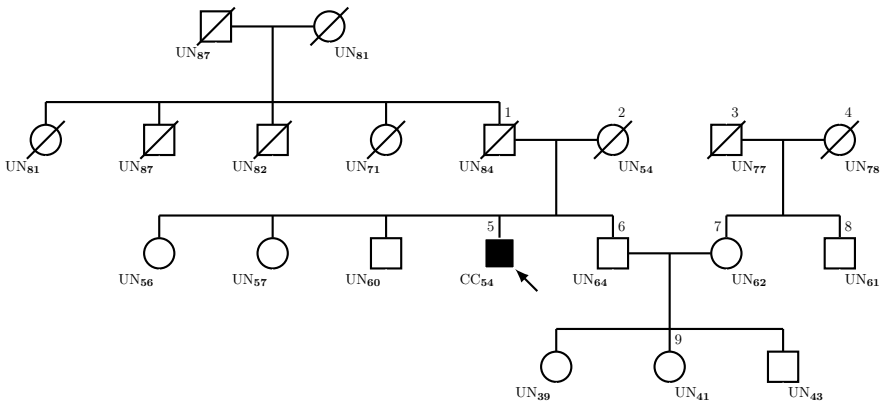
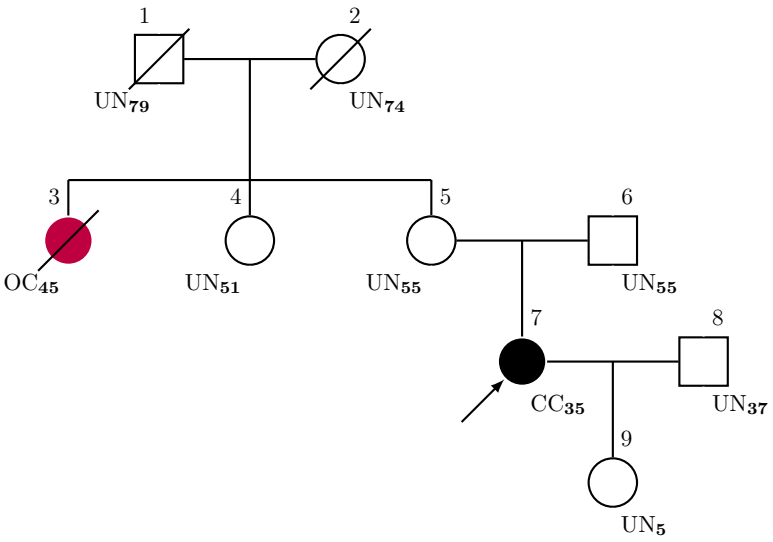


Figure 8.10: Large family of unaffected individuals and one case of CC diagnosed at age 54.

ted a deleterious variant to his daughters. Furthermore, the removal of partial data against LS leads to increased (respectively decreased) risks for all his blood relatives, i.e. Individuals 3, 4, 5, 7 and 9 (respectively non-blood relatives, i.e. Individuals 2, 6 and 8). We put emphasis again on the importance of considering the familial structure by comparing individual posterior probabilities conditional on associated individual phenotype only (first line) and conditional on the whole FH (second or third line). The accumulation of cases greatly increases posterior probabilities of carrying LS when conditioning on the FH for all blood relatives of the cases. The only individual exempted of an increased posterior probability of LS is Individual 8 who is no blood relative of any family member except Individual 9 whose phenotype is nearly uninformative (UN_5). The two cases have different posterior LS probability conditional on their phenotype du to different penetrance of MMR genes per localization (in particular lower penetrance for ovarian cancer) and different ages at diagnoses. However their posterior probabilities are almost equal when conditioning on the FH du to the family structure.

Moreover, protective phenotypes of Individuals 4 and 5 lead to a posterior LS probability conditional on their respective phenotype lower than the general population (0.35) but the posterior LS probability for individual 4 (respectively 5) is about half (respectively equal) the one of the cases. Indeed, if a deleterious allele is carrier by Individual 7 it is nearly certainly transmitted by her mother. Posterior risks for Individual 4 are about half her sisters, only slightly lowered by her protective phenotype. Similarly posterior probabilities of carrying LS for Individual 1 and 2 conditional on their phenotype is below those of general population du to their protective phenotype (UN_{79} and UN_{74} respectively) and rises when conditioning on FH in lower proportions than Individual 5 as they share the risk of having transmitted a deleterious variant to their daughters. As previously mentioned, the posterior probability of carrying LS for Individual 1 is higher than the one of Individual 2 as LS carrier females are less likely to survive free of disease up to a great age du to the presence of EC and OC in the Lynch spectrum. Finally Individual 9 having a nearly uninformative phenotype (UN_5) her posterior probability of carrying LS conditional



(a) Family history FH9

$\mathbb{P}(X_i \neq (0000)\|\text{ev})$	i=1	i=2	i=3	i=4	i=5	i=6	i=7	i=8	i=9
$\text{ev} = \{\text{PH}_i = \text{ph}_i\}$	0.21	0.16	6.52	0.31	0.29	0.32	18.87	0.35	0.36
$\text{ev} = \{\text{FH9}\}$	10.36	8.93	18.63	7.60	18.50	5.80	23.86	0.35	12.10
$\text{ev} = \{\text{FH9}_{\setminus 1}, \text{PH}_1 = \text{UN}_{28}\}$	31.30	6.89	37.11	14.81	36.93	4.58	40.93	0.35	20.63

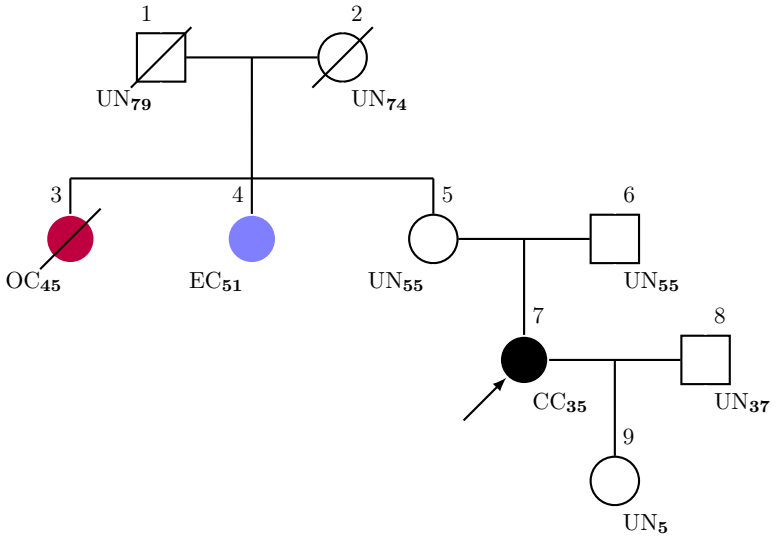
Figure 8.11: An example of a family denoted FH9 and computed LS probabilities (in percent) conditional on FH9 in the hypothetical situation where Individual 1 were either dead at age 79 (second line) or dead at age 28 (third line) free of disease. We denote by $\text{FH}_{\setminus i}$ the set of phenotypes in the family except the phenotype of Individual i.

on her phenotype is almost equal the one in the general population (0.35) and rises to nearly half her mother's one when conditioning on FH.

Let us now propose an overview of computed future disease risks conditional on chosen examples of FH. We firstly recall that, as previously mentioned in Section 8.2.2.3, we ignore the competing risk with death. As death precludes any other event to happen, risks of diseases are prone to be overestimated in LynchRisk, in particular when ignoring the (duration dependent) excess of risk of death after disease. This will be corrected in future versions however, a semi-Markov model requires more challenging calculations. Note that competing risk with death is also ignored in MMRpro however, as the authors assume a competing risk model with no transition between diseased states. In such models, no excess of risk of death after disease is involved and therefore the error made is probably smaller, at the cost of ignoring multiple cancers affecting the same individual.

We start with a sever FH of cancer denoted FH10 and pictured in Figure 8.12a. Figure 8.12b highlights the importance of taking into account the FH for computing future disease risks of patient 5 as well as the evolution of these risks while the patient remains free of disease through time. The left column reports the evolution of her posterior LS probability conditional on $\text{FH10}_{\setminus 5}$ and her phenotype while she remains free of disease at age $s = 60$, 65, 70 and 75. We assume that $\text{FH10}_{\setminus 5} = \{\text{PH}_1, \ldots, \text{PH}_9\} \setminus \text{PH}_5$ is fixed, hence data for other family members are independent of time and fixed at associated phenotypes in FH10. With no surprise, the posterior probability of LS for Individual 5 decreases with time as she remains free of cancer which impact her posterior probability of future cancer development as reported in the four middle columns. The weight of the family history in posterior disease risks calculation is highlighted in the four middle and four right columns reporting respectively, her posterior disease risk at age t conditional on $\{\text{FH10}_{\setminus 5}, \text{PH}_5 = \text{UN}_s\}$ while she remains free of disease at age s and the ratio between these probabilities and the probabilities of disease risks conditional on $\text{PH}_5 = \text{UN}_s$ only. Quantities are computed for various times t and s. The weight of FH10 with several severe cases lead to relative risks varying from 9.44 to 19.78. However relative risks decrease with time in both time scales.

We have previously seen certain types of posterior familial risks such that the posterior distribution of the number of carriers in a family. LynchRisk computes other types of familial risks which could be valuable information along with individual risks in order to help clinicians at identifying families at risk. For instance, LynchRisk computes the posterior probability of observing at least one cancer in any at risk (i.e. alive) individual in a family by a chosen time. Results for a selection of two families (FH9 and FH10) within 5, 10 and 15 years are reported in Table 8.9. Note that both families are the same except Individual 4 who is either unaffected or diseased with EC at age 51. They contain seven at risk, i.e. alive, individuals. As expected the posterior probability of diagnosing at least one cancer within any period of time increases with the severity of the history of disease in the family. We believe that such quantities are valuable additional information for risks evaluation. In particular for instance, the intuitive concern about the extreme severity of FH10 is reinforced by a 41% risk of observing at least one cancer in the family within 5 years and nearly



(a) Family history FH10

	Post. LS probability	Post. disease risk before age t t = 60	65	70	75	Relative risk t = 60	65	70	75
$s = 55$	92.53	16.5	33.41	45.72	57.75	19.78	16.37	12.32	9.90
$s = 60$	91.12	0.00	20.25	35.00	49.40	–	16.64	12.06	9.80
$s = 65$	89.00	–	0.00	18.50	36.55	–	–	10.84	9.44
$s = 70$	86.73	–	–	0.00	22.15	–	–	–	10.06
$s = 75$	83.33	–	–	–	0.00	–	–	–	–

(b) Various posterior probabilities for Individual 5 conditional on FH10

Figure 8.12: An exemple of a FH denoted FH10 and various quantities associated with Individual 5: 1) Posterior LS probability, in percent, conditional of FH10 and $\text{PH}_5 = \text{UN}_s$ for $s = 55, 60, 65, 70, 75$ (left column), 2) Posterior probability (in percent) of being diagnosed with a Lynch-associated cancer before age t conditional on FH10 and $\text{PH}_5 = \text{UN}_s$ (middle columns), 3) Ratio between her posterior probability of being diagnosed with a Lynch-associated cancer before age t respectively conditional on FH10 and $\text{PH}_5 = \text{UN}_s$ and conditional on $\text{PH}_5 = \text{UN}_s$ regardless FH10 (right columns).

	5 years	10 years	15 years	20 years
FH9	9.33	17.79	26.36	35.25
FH10	40.99	64.41	79.82	89.05

Table 8.9: Probability (in percent) of diagnosing at least one Lynch-associated cancer within 5, 10, 15 or 20 years among the 7 persons at risk (alive) in families FH9 and FH10.

90% within 20 years.

We finally extend aforementioned posterior disease probabilities to chosen subsets of localizations. As seen in section 2.3, one can select a subset of states in LynchRisk's multi-state model in order to compute posterior disease risks in a selection of localizations. Table 8.10 reports posterior future disease risks for Individual 7 and 9 in FH10 in the absence of risk reduction surgery (hysterectomy and/or bilateral salpingo-oophorectomy). For the sake of simplicity, we assume that biological testing results are absent as well as germline screening testing (which may confirm LS or be inconclusive). Probabilities are also computed with MMRpro (Mp) if applicable, i.e. for individuals free of disease (Individual 9) as MMRpro's model is a competing risk model considering the first diagnosed disease only and for aggregated localizations colon/rectum as well as endometrium. We must stress that the tendency of higher risks predicted by LynchRisk for Individual 9 when compared to MMRpro are partly be explained by a higher posterior LS probability computed for her (46.3% by LynchRisk versus 20.7% by MMRpro) mostly explained by Individual 3 diagnosed with an ovarian cancer at age 45 being considered as free of disease at 45 in MMRpro's calculations.

Note that the predicted risk of gynaecological cancer is the sum of future risk of endometrial cancer and ovarian cancer for Individual 7 but slightly below that sum for Individual 9. Indeed, as there is only one path from a diseased state $D \in \mathcal{D}$ to a subsequent diseased sate $\tilde{D}$ for any disease $\tilde{D} \in \mathcal{D}$, future risks in groups of localizations can simply be added for a diseased individual. However several paths being possible from state "UN" to state $\tilde{D}$ renders computations more challenging for unaffected individuals where one path, for instance $\mathrm{UN} \to \mathrm{E} \to \tilde{\mathrm{O}}$ participates once in risks of gynecological cancer and also once for each future risks of endometrial cancer and ovarian cancer separately.

Posterior probabilities of carrying LS for Individual 7 and Individual 9 computed with LynchRisk are as high as 92.5% and 46.5% respectively. These results are explained by the severity of the FH with three sever cases rendering Individual 7 nearly certainly carrier. With a nearly non-informative phenotype (unaffected at age 5), the posterior probability for Individual 9 is about half her mother's one. Predicted disease risks are consistent with a high posterior probability of carrying LS for each individual respectively. In particular for instance the risk of colon, endometrial and ovarian cancer for Individual 7 reaches respectively 23.14%, 26.23% and 7.70% by age 70. Usually risk reduction surgery is proposed to patients of high probability of carrying LS according to their age, parenting project and psychological impact. The posterior probability of carrying LS for Individual 9 is about half her

Loc. \ Age	40	50	60	70
Colon	4.63	10.44	16.99	23.14
Rectum	0.18	0.90	2.02	3.35
Endometrium	1.30	10.11	22.10	26.23
Ovary	0.91	4.25	7.05	7.70
Upper GI	0.17	1.15	2.39	3.59
Urinary tract	0.28	1.20	2.60	4.23
Gynaecological	2.21	14.36	29.15	33.93
Any	7.46	28.05	53.16	68.23

(a) Individual 7 ($PH_7 = CC_{35}$)

Loc. \ Age	LR 30	Mp 30	LR 40	Mp 40	LR 50	Mp 50	LR 60	Mp 60	LR 70	Mp 70
Colon	0.25	0.34	4.06	1.32	7.28	4.18	11.91	6.54	17.17	8.07
Rectum	0.01		0.11		0.59		1.57		3.04	
Endometrium	0.01	0.03	0.80	0.27	5.87	2.09	13.66	7.01	17.23	9.71
Ovary	0.03		0.80		2.79		5.00		5.78	
Upper GI	0.01		0.16		0.80		1.89		3.37	
Urinary tract	0.02		0.25		0.85		2.08		3.87	
Gynaecological	0.03		1.59		8.36		17.62		21.72	
At least one	0.33		6.00		16.00		28.61		37.21	
At least two	0.00		0.41		3.16		10.80		18.79	

(b) Individual 9 ($PH_9 = UN_5$)

Table 8.10: Posterior future cancer risks for Individual 7 (on the top) and Individual 9 (at the bottom) conditional on FH9 computed with LynchRisk (LR) and MMRpro (Mp) if applicable. GI stands for gastrointestinal and gynaecological aggregates endometrial and ovarian cancer.

mother's one leading to predicted risk per localization ranging from 3.09% to 17.39% and reaching 43% for at least one cancer by age 70.

8.5 Discussion and perspectives

8.5.1 Clinical context

LS carriers have a very high risk of developing a CRC (52% by age 70) and an EC (45% by age 70) and increased risks in other localizations (e.g. stomach, small bowel, ovary, urinary tract, urinary bladder, bill duct, gallbladder, pancreas, etc.). Their risk of developing any Lynch-associated cancer rises up to 70% for males and nearly 80% for females by age 70 (Dominguez-Valentin et al., 2020b). Therefore LS detection is crucial for adapting surveillance of patients and their family members. However

germline screening all individuals affected by a Lynch-associated cancer would be unrealistic for healthcare structures and often inconclusive with VUS screened or lack of sensitivity of screening technics. Moreover, the psychological impact can be important. In this context, the family history of a patient is one of an essential tools for assessing the probability of carrying LS.

Additionally, current guidelines now recommend universal screening of MSI status for all CRC or EC before age 70 (Vasen et al., 2013). Universal screenings means MSI status detection by PCR-based technics or NGS and/or IHC for the four MMR proteins on a tumoral tissue. Such recommendation are advocated both because an MSI status is highly associated with LS and because MSI tumors show different responses to treatments, hence they are candidate for personalized medicine. In particular MSI tumors show strong responses to immune checkpoint inhibitors at a metastatic stage. Currently, there is no recommendation of MSI screening for extra-colorectal, extra-endometrial tumors. Most MSI tumors are however sporadic (i.e. not due to LS). Additional biological testing negatively associated with LS such as MLH1 promoter hypermethylation testing or BRAF V600E mutation screening are recommended in tumors showing lack of MLH1/PMS2 expression. Nevertheless, none of the above data can confirm or rule out LS with certainty and the need of mathematical tools which integrate data along their availability is increasingly needed for helping clinicians in their decision making.

8.5.2 Current mathematical models

There exists currently two main mathematical models for assessing probabilities of carrying LS called PREMM_5 and MMRpro. They both have being presented in Section 7.3.3 and showed good predictive performances for triaging individuals according to their probability of carrying LS on their respective validation dataset. They are rarely compared, however Win et al. (2013) performed a meta-analysis of 17 studies and report AUCs ranging from 0.80 to 0.84 for triaging LS versus non-LS individuals. Mathematical models are however poorly implemented in daily clinical practice and they are not included in current guidelines except those of the NCCN[8] who advocates the use of PREMM_5. We believe that despite their good performances, they show some limitations in regard of the evolution of the availability of biological data and the inclusion of new localizations in the Lynch spectrum in the past decades.

PREMM_5 is a logistic-regression based model implemented in a user-friendly web interface[9]. Unlike pedigree-based models, it solely takes into account partial information on the proband and on affected first and second degree relatives. Therefore, data entrance is easier and faster but important information is ignored in the calculation. In particular, the structure dependency between genotypes of family members is partially ignored, the number and parental relationships of unaffected individuals is not taken into account and ages at diagnoses are ignored in some (less relevant) situations. CC and RC are aggregated into CRC, despite the fact that MMR genes are more penetrant for CC than RC. Moreover, PREMM_5 is not designed for in-

[8] `https://www.nccn.org`

[9] https://premm.dfci.harvard.edu

tegrating biological testing results despite their high association with LS and their inclusion into the Bethesda guidelines (Rodriguez-Bigas et al., 1997). These data become increasingly available with universal MSI screening recommended for all CRC and EC before age 70.

MMRpro is a pedigree-based model and therefore it takes into account the full dependency structure between genotypes of family members and the entire set of time-to-event data. Additionally it offers a predicting value for future cancer risks. However its parameters have not been updated since its development in 2006. Moreover CC and RC are aggregated into CRC and no other Lynch associated cancer than CRC and EC is considered. However the Lynch spectrum is wide and considering extra-colonic, extra-endometrial cancers is advocated since the extension of the Amsterdam criteria to the Amsterdam criteria II (Vasen et al., 1999). Moreover, subsequent cancers are ignored whereas LS carriers are fairly often prone to developing more than one cancer in a lifetime. Finally, some biological testing data are also ignored, for instance the information regarding which protein shows lack of expression by IHC is not taken into account and neither MLH1 promoter hypermethylation nor BRAF V600E mutation is included in the model.

We believe that an updated mathematical model that overcomes these limitations would be an important tool for clinicians in their assessment of posterior probabilities of carrying LS and future cancer risks predictions. For that purpose we implemented a pedigree-based model called *LynchRisk* which takes into account the main four MMR genes (MLH1, MSH2, MSH6, PMS2), cancer localizations colon, rectum, endometrium, ovary, upper gastrointestinal tract (which aggregates stomach, small bowel, bile duct, gallbladder and pancreas) and urinary tract (which aggregates ureter, kidney and urinary bladder). LynchRisk considers up to two cancers per individuals and computes a posterior probability of carrying LS (overall and per gene) for any family member as well as future cancer risks per localization or an overall risk for any unaffected individual or affected individual with at most one cancer. Biological data are integrated with MSI status, IHC testing result per protein, MLH1 promoter hypermethylation and BRAF V600E screening. Note that an important advantage of logistic-regression based models over pedigree-based models in daily clinical practice is their simplicity and rapidity of use as FH data are only partly entered. We therefore believe that LynchRisk could be either an important tool by itself or complementing a logistic-regression such as PREMM_5 for assessing risks when results are complicated to interpret. However, the importance of collecting data laying in entire FHs become more and more highlighted both for risks assessment and epidemiological studies and therefore FHs are now more and more entirely considered and reported.

First computed posterior probabilities on a variety of simulated datasets show coherent results with expected values. We saw the importance of considering both the set of available biological testing results increasingly available and a wide spectrum of localizations and multiple onsets affecting an individual in risks assessments. Simulated examples highlighted the importance LynchRisk specificities versus PREMM_5 and MMRpro. Future cancer risks may however be overestimated due to the non consideration of competing risks with death. A lot of exciting and challenging work

ahead is still needed.

8.5.3 Validation

Our first future objective is the validation of LynchRisk and the choice of a threshold over a clinic-based cohort in collaboration with clinicians and biologists including Patrick Benusiglio (Hôpital universitaire Pitié Salpêtrière, APHP, Sorbonne Université / CRSA, Hôpital Saint-Antoine, Paris), Alex Duval (CRSA, Hôpital Saint-Antoine, Paris), Florence Coulet (Hôpital universitaire Pitié Salpêtrière, APHP, Sorbonne Université / CRSA, Hôpital Saint-Antoine, Paris) and Erell Guillerm (Hôpital universitaire Pitié Salpêtrière, APHP, Sorbonne Université, Paris). Saint-Antoine hospital hosts one of the largest clinic-based cohort of CRC patients with more than 2000 patients including 300 confirmed LS and an estimate of 80 Lynch-like patients from the Saint-Antoine hospital and universities and regional hospitals around Paris. Tumoral MMR protein expression analyses by IHC and promoter hypermethylation are performed on site at Saint-Antoine hospital and MSI testing is done with Pentaplex or Next Generation Sequencing (NGS) if needed at the partner laboratory in la Pitié Salpêtrière hospital. Model validation will be done in collaboration with the clinical cancer genetics team which has a sustained activity in genetic counseling and surveillance of patients. Family histories, clinical and molecular data are entered prospectively since 2017 in an electronic database. Older reports are still mostly in paper files but regularly entered into the electronic database. We plan to start the model validation once all reports are fed into the database. However reports are numerous and their entrance require good clinical knowledge.

8.5.4 Parameter estimation

The main drawback of LynchRisk is its parameters which are extracted and/or estimated from various sources in the literature. We chose estimates of allele frequencies reported by Win et al. (2017) as the most recent and reliable resource. The method used by the authors is fairly well explained however the authors omit important details such that, for instance, details about the survival model used (parametric or not). Furthermore, LS prevalence in the general population is prone to controversies with discrepant estimates up to five fold which suggests that a thorough meta-analysis should be performed.

We are quite confident with estimates of sensitivity and specificity of biological tests on CRC reported by Assasi et al. (2016). The authors performed a large meta-analysis with clear exclusion criteria. Note however that the study was published in 2016 and sensitivity and specificity may vary with time. Moreover NGS sequencing may replace or complement PCR-based methods in the future. Finally, colon and rectal cancer are pooled although MSI status in each localization may have different predictive values for LS[10]. Note also that the probability of MLH1 promoter hypermethylation is increasing with the age[11] although we assume it to be constant with respect to time.

[10] https://www.insight-group.org/syndromes/lynch-syndrome/
[11] https://www.insight-group.org/syndromes/lynch-syndrome/

Estimates of sensitivity and specificity of biological tests in EC are scarce, not to say absent, in the literature and we assumed them to be equal those in CRC. Clearly that strong assumption is not completely reasonable. In particular MSH6 is highly penetrant for EC and more often associated with MSI-L (MSI-low) tumors than MSI-H (MSI-high) tumors suggesting a probable lower sensitivity of MSI testing in EC than in CRC.

We firstly planed to include the histological profile of ovarian carcinoma in LynchRisk. However estimates of sensitivity and specificity of such profile is nearly absent with only few and very small studies conducted towards that goal. If more data become available, a meta-analysis will be needed. We decided to exclude that variable in LynchRisk's first version.

Finally, last but not least, the width of the Lynch spectrum and the frequency of LS carriers diagnosed with more than one cancer led us to choose the survival model composed of multiple transient states represented in Figure 8.1. We chose a piecewise constant hazard model for its clear interpretation and because it leads to closed-form formulae for computing transition probabilities in the discretized Markov model developed in Section 2.3. With a model composed of such a large number of transitions, each of them being sex and genotype dependent, we faced the problem of estimates availability in the literature. We therefore had to make numerous assumptions listed in Section 8.3.3. In particular we assumed that diseases occur independently, incidences of a subsequent cancer are independent of the past history of disease and effects of MMR variants are additive on the hazard. Note however that strong assumptions are commonly made in genetic models, prone to similar data restrictions.

Despite these assumptions and the important subsequent dimension-reduction, we still faced a problem of overfitting issues for carriers of a pathogenic MMR variant. Working on time-dependent hazard ratios, rather than hazards themselves, between carriers and non-carriers makes much more sense in genetic diseases as the core question is a relative risk between carriers and non carriers. We assumed that hazard ratios are piecewise constant. In the absence of available data for non-carriers, we focused on hazard ratios between carriers and the French general population and derived hazard rates for non-carriers from the mixture form of the hazard function in survival analysis. However, estimates reported in the literature are pooled per time-intervals. In the absence of individual data, no cross validation was possible for performing hazard ratio regularization and we finally performed a model selection based on a likelihood ratio test.

To recap, we faced several issues for survival parameter estimation in LynchRisk and finally obtained satisfying estimates but far from being fully satisfying. Therefore hazard ratio estimations is our second main goal for future versions of LynchRisk. A polygenic effect and frailty models will also be considered in future versions. Note that an important and interesting point in hazard ratio estimation in familial genetics is the ascertainment bias as families are seen conditional on the fact that the proband came for investigation. Hence families seen in genetic counseling are "selected" by severity of diseases and the frequency of pathogenic alleles within these families are higher than in the general population. A current and common method for correcting

ascertainment bias consists in ignoring the phenotype of the proband (the person who seeks medical attention). The underlying idea is that the proband is the reason why the family is seen in consultation. However, one can intuitively understand that such method is not fully satisfying. It often tends to remove an individual with an early onset and results in underestimation of incidences at young ages.

8.5.5 Additional future extensions

We also plan on including other features in the model, starting with the inclusion of the CMMRD syndrome. The CMMRD syndrome is defined as biallelic germline mutations in the same MMR gene and characterized by early-onset and often multiple CRC, lymphomas/leukemias, brain tumors in childhood (Lavoine et al., 2015; Buecher et al., 2019). Ignoring the CMMRD syndrome in the first version of LynchRisk is acceptable as this syndrome is rare and associated with a sombre prognosis. Thus CMMRD patients rarely have descendants. However as LynchRisk models the transmission of alleles between family members, including homozygous carriers is important and it should be done in future versions of LynchRisk in collaboration with Alex Duval, Patrick Benusiglio and Chrystelle Colas. However the rarity of the syndrome leads to scarce incidence data per CMMRD-associated cancer type. The associated phenotypes may be added as a binary variable as a start.

The Markov assumption is commonly made in multi-state survival as semi-Markov models lead to more challenging computations and estimations of transition probabilities. However the Markov assumption implies that we ignore in particular the duration dependent excess of risk of death after disease which leads to no bias to compute posterior probabilities of carrying LS under the assumption such that the hazard of the excess of risk of death is additive and independent of the genotype. However that assumption is untrue as LS patients show better prognosis than non-LS patients. Furthermore, ignoring the excess of risk of death leads to increased posterior future cancer risks. A semi-Markov model will be considered in a more distant future.

8.5.6 Variants of uncertain significance

Let us finally conclude with a final remark regarding another potential usefulness of LynchRisk. We hope and believe that LynchRisk can not only be a tool for clinicians in their risk assessment and decision making but also a valuable tool for committees in charge of classifying MMR variants. About a third of germline screened variants are of unknown or uncertain significance (VUS) leaving clinicians, patients and geneticists with uncertain conclusions. With the increasing use of NGS, the number of sequenced VUS is growing since the last decades. The InSiGHT database[12] along with the ClinVar database[13] is one of the largest database of MMR variants classified in five-tiered categorical scale: “pathogenic” (Class 5), “likely pathogenic” (Class 4), “uncertain” (Class 3), “likely not pathogenic” (Class 2) or “not pathogenic” (Class 1). A scientific committee aims at classifying high penetrant from low penetrant variants

[12] https://www.insight-group.org/variants/databases/
[13] https://www.ncbi.nlm.nih.gov/clinvar/

(no middle risk variant is considered), hence refining classification towards Class 1 or Class 5 as evidence data become available. Their method detailed in (Goldgar et al., 2004, 2008; Easton et al., 2007; Thompson et al., 2013, 2014) is based on a likelihood ratio testing the hypothesis of a pathogenic variant versus a neutral variant using a variety of evidence data including in silico data (sequence conservation and position as neutral variants are more likely to be conserved across generations), in vitro data (results of functional assays), clinical evidences (co-occurence of variants as homozygous carriers of pathogenic variants are unlikely, unless if associated with a CMMRD phenotype) and segregation analyses. We assumed so far that genotypes are latent variables and we focused on computing posterior LS carrier probabilities conditional on an evidence. However one can compute with the same algorithm $\mathbb{P}(\text{ev}|X_I^g = x_I^g)$ for any gene G^g where $X_I \subseteq \{X_1, \ldots, X_n\}$ is a subset of sequenced genotypes in the family and x_I^g the set of observations. Hence the posterior probability of observing a family history and various biological testing conditional on one or a set of observed (sequenced) genotypes could be an additional tool to be included in the likelihood ratio used for variants classification.

General conclusion

Summary. This thesis deals with exact inference in probabilistic graphical models and in particular in Hidden Markov Models (HMMs) and Bayesian Networks (BNs). It provides extensions and various applications of the sum-product algorithm (also called belief propagation in BNs and forward-backward in HMMs) in the fields of segmentation, multi-state survival and familial genetics. It has been driven by the goal of developing a pedigree-based model for computing probabilities of genetic predisposition and cancer risks in the framework of the Lynch syndrome. Part I provides an introduction to the sum-product algorithm (Chapter 1) and survival analysis (Chapter 2), the main two algorithmic and statistical tools developed and/or applied throughout the thesis. It also includes basics in molecular biology and genetics (Chapter 3) needed for understanding the notion of genetic predisposition and pedigree-based models.

My thesis started in Part II with a reflection about possible extensions of the work of Cowell (1992); Nilsson (2001); Aston and Martin (2007); Nuel (2019) for exploring novel uses of polynomial potentials and generating functions in the sum-product algorithm. This part led to two main contributions. We firstly proposed, in Chapter 5, a method based on generating functions of the derivatives of local probability distributions and an adaptation of the classical product of polynomials for computing the exact derivatives of the likelihood in a BN up to a chosen order d. The complexity of our method is in $\mathcal{O}(c \times d^2)$ in time for a uni-dimensional parameter and $\mathcal{O}(c \times p^{2d})$ for a p-dimensional parameter, $p > 1$, where c is the time complexity for computing the likelihood. We secondly expose how a segment-based HMM can be used for allowing an arbitrary prior on the number of segments in the field of sequence segmentation and we show that recursions over polynomial potentials lead to a compact implementation and a complexity reduction. Both methods were in particular applied to the constrained segment-based HMM introduced in Mercier and Nuel (2021) in order to offer an unsupervised method for the statistical learning of a scoring function for the maximal scoring segment and extend the maximal scoring segment to multiple segments and multiple segment types (Chapter 6).

We pursue in Part III with the development of a pedigree-based model called LynchRisk for computing individual probabilities of genetic predisposition and future cancer risks conditional on a family history of cancer and various biological and clinical test results in the framework of the Lynch Syndrome (LS). Chapter 7 is an introductory chapter which describes the epidemiological, biological and clinical context the model was built in. Due to the wide spectrum of LS and the frequency

of multiple cancers in LS carriers, a multi-state survival model with several transient states has been chosen for modeling personal histories of cancer. We explain in Chapter 2 how such model can be implemented in a discretized HMM in order to avoid the formal computation of exponential matrices needed for solving Kolmogorov forward differential equations. The description of the model and its parameters, extracted or estimated from exhaustive statistics available in the literature, is provided in Chapter 8 which combines tools and notions seen in Chapters 1, 2, 3, 4 and 7. Chapter 8 also presents an extensive comparison over simulated datasets of results obtained with LynchRisk and with PREMM_5 and MMRpro, the main two current statistical models for computing probabilities of carrying LS and/or developing cancers of the Lynch spectrum.

Perspectives. These first steps open up several questions and perspectives in different fields. A comparison of various methods used for computing the score and observed Fisher information matrix is essential for the work started in Chapter 5. We firstly would like to compare smoothing recursions proposed by Cappé and Moulines (2005) for computing Louis identities (Louis, 1982) with the one developed by Cowell (1992) and Nilsson (2001) for the computation of order two moments of decomposable functions. We secondly would like to apply our method for the computation of the order one derivative in Oakes *formulae* (Oakes, 1999) for proposing an extension of Oakes' work to a general BN setting. That latter approach seems promising in terms of easy implementation and computational complexity reduction. We also should compare these various methods with sensitivity analysis and automatic differentiation.

Chapter 6 seems to offer an interesting approach for the unsupervised learning of a scoring function for the maximal scoring segment with extensions to multiple segments and multiple segment types. To the best of our knowledge, current scoring functions are empirically learnt according to the feature of interest in target segments. However our method shows a lower sensitivity in a particular framework of a well known feature of interest (for instance transmembrane amino-acid properties to localize a transmembrane protein domain). A main perspective of this work is its application in a field in need of unsupervised methods when the feature of interest is not or insufficiently known. Secondly, whereas the interest of extending the maximal scoring segment to multiple segments seems straightforward, it is still unclear whether a segment-based approach offers a statistical contribution in comparison to classical HMM or not. It may indeed be irrelevant to control the prior on the number of segments in segment-based HMMs versus the prior on the transition probabilities in classical HMMs. We are still unsure of the answer to that question which constitute another main perspective of this work. Nevertheless a segment-based model as the one described in Section 6.3 allows one for mixing different probability distributions conditional on the number of segments and may open up interesting perspectives that we would like to explore.

The first version of LynchRisk detailed in Chapter 8 is promising when compared to the main two other models over simulated datasets. However a validation step on a real dataset still needs to be performed before publication and constitute the

main perspective for this work. We plan on using the extensive database of Saint-Antoine hospital whose digitization is still ongoing. Several updates are next on the list of perspectives and in particular hazard rates estimation. One can also consider a polygenic effect and/or a frailty model still absent in the first version of LynchRisk. We also would like to include the CMMRD syndrome but we face the problem of insufficient data for estimating hazard rates per CMMRD localization for carriers of that syndrome. Its inclusion, first, as a binary variable could be a primary option we are working on. Note also that LS *de novo* mutations are ignored in the first version of LynchRisk and their inclusion could constitute another perspective. We expect to postpone this perspective to a more distant future as *de novo* mutations in the framework of LS seem very rare which involves, firstly, several issues for estimating their frequency (still prone to controversies) and secondly, law bias in risks computations. Finally, modeling personal histories of cancer in a semi-Markov model could also be an interesting perspective for taking into account the excess of risk of death after cancer. We however decided to leave it also to a more distant goal for it to induce much more complexe computations whereas the Markov property induces no bias in posterior probabilities of carrying LS if the excess of risk of death is assumed to be additive on the hazard. Moreover the Markov property induces a conservative bias on future cancer risks as such property increases the probability of being at risk, that is alive, for an individual diagnosed with one cancer in our model.

Finally we are working on an R package for pedigree analysis. We propose a pedigree-based model adaptable to a variety of diseases involving a genetic component. Outputs will include individual posterior probabilities of carrying deleterious variant(s), the posterior distribution of the number of carriers, posterior carrier probabilities conditional on that number and, provided additional input data related to the genetic model, future disease risks.

Appendix A

Essential definitions in graph theory

Sommaire

This brief appendix lists essential definitions in graph theory.

Definition 12 (graph)**.** *A graph $G = (X, \mathcal{E})$ is a set of nodes X and a set of edges $\mathcal{E}$.*

In this document, we assume X to be a finite set of indexed random variables $X = \{X_1, \ldots, X_n\}$.

Definition 13 (directed edge, undirected edge, parent, child, neighbor)**.** *A directed edge is an ordered pair $(X_i, X_j) \in \{X_1, \ldots, X_n\}^2$ and is symbolized by an arrow. If $X_i \to X_j$ we say that X_j is a child of X_i and X_i is a parent of X_j. An undirected edge is symbolized by a bar $X_i - X_j$ and is equivalent to $X_i \rightleftarrows X_j$. If $X_i - X_j$ we say that X_i and X_j are neighbors. A directed (respectively undirected) graph is a graph containing only directed (respectively only undirected) edges. A mixed graph is a graph containing directed and undirected edges.*

Definition 14 (induced subgraph)**.** *Let $G = (X, \mathcal{E})$ be a graph and $S \subseteq X$, the subgraph induced by S is the graph $G[S] = (S, \mathcal{E}')$ where $\mathcal{E}' = \{(X_i, X_j) \in \mathcal{E}; X_i, X_j \in S\}$.*

Definition 15 (clique)**.** *Let $G = (X, \mathcal{E})$ be an undirected graph and $C \subseteq X$, the subgraph $G[C]$ induced by C is complete if every pair of nodes in $G[C]$ are connected by an edge. The set of nodes C of a complete subgraph $G[C]$ is called a clique of G.*

Definition 16 (maximal clique)**.** *A clique in an undirected graph is said to be maximal if the addition of any node renders it not to be a clique.*

Definition 17 (simplicial node). *A node is simplicial in a graph if its neighbors form a clique.*

Definition 18 (topological ordering). *Let $G = (X = \{X_1, \ldots, X_n\}, \mathcal{E})$ be a directed graph, we say that $X_1, \ldots, X_n$ are in topological ordering in G if whenever we have $X_i \to X_j$, then $i < j$.*

Definition 19 (trail). *A trail in a graph $G = (X, \mathcal{E})$ is a sequence of nodes and edges $(X_0, \mathcal{E}_1, X_1, \mathcal{E}_2, \ldots, X_{k-1}\mathcal{E}_k, X_k)$ such that, for all $i \in \{1, \ldots, k\}$, $\mathcal{E}_i = (X_{i-1}, X_i)$ and all edges are distincts.*

Definition 20 (path). *A path is a trail in which all nodes are distincts.*

Definition 21 (descendant). *Let $G = (X = \{X_1, \ldots, X_n\}, \mathcal{E})$ be a graph, we say that X_j is a descendant of X_i if there exists a directed path from X_i to X_j.*

Definition 22 (connected graph). *A graph is connected if there exists a path between each pair of its nodes.*

Definition 23 (loop). *A loop is an edge connecting a node to itself.*

Definition 24 (leaf). *A leaf in an undirected graph is a node connected to exactly one other node by an edge. A leaf in a directed graph is node with no child.*

Definition 25 (cycle). *A cycle is a trail $(X_k, \ldots, X_\ell)$ with $k = \ell$*

Definition 26 (directed acyclic graph). *A directed acyclic graph (DAG) is, as its name speaks for itself, a directed graph with no cycle.*

Definition 27 (chord). *A chord in a cycle is an edge connecting two non adjacent nodes.*

Definition 28 (chordal graph). *A chordal graph is a graph in which each cycle of four or more nodes has a chord.*

Definition 29 (tree). *A tree is a connected undirected graph with no cycle and no loop.*

Definition 30 (probabilistic graphical model). *A probabilistic graphical model is a graph whose nodes are random variables or sets of random variables and edges express probabilistic relationships between variables.*

Appendix B

Computing Individual Risks based on Family History in Genetic Diseases in the Presence of Competing Risks

Sommaire

This appendix contains an article published in *Computational and Mathematical Methods in Medicine* in 2017, at the end of my Master's thesis. This work is a joint work with my PhD supervisor, Grégory Nuel (CNRS, LPSM, Sorbonne Université, Paris), and Olivier Bouaziz (Université Paris Descartes, Paris). During my Master's thesis, I implemented in `R` the Claus-Easton model (Claus et al., 1991; Easton et al., 1993) for it to be used by the Institut Curie[1] as a primary tool for assessing risks of genetic predisposition in the framework of the breast/ovarian syndrome. This work was done in collaboration with Antoine de Pauw (Institut Curie, France). The Claus-Easton model was later used in the following article for illustration. This article details essential tools for computing risks in the framework of diseases with a genetic component and a particular emphasis is put on a competing risk setting with death. Its main contribution is related to the expression of the time-dependent posterior hazard rate in that framework.

[1] `https://curie.fr`

Research Article: Computing Individual Risks based on Family History in Genetic Disease in the Presence of Competing Risks

G. Nuel*[1,2], A. Lefebvre[3,4] and O. Bouaziz[5]

[1]LPMA, UMR CNRS 7599, Paris, France
[2]UPMC, Sorbonne universités, Paris, France
[3]UPSud, Paris-Saclay, Orsay, France
[4]Institut Curie, Paris, France
[5]MAP5, UMR CNRS 8145, Paris, France

September 13, 2017

Abstract

When considering a genetic disease with variable age at onset (ex: diabetes, familial amyloid neuropathy, cancers, etc.), computing the individual risk of the disease based on family history (FH) is of critical interest both for clinicians and patients. Such a risk is very challenging to compute because: 1) the genotype X of the individual of interest is in general unknown; 2) the posterior distribution $\mathbb{P}(X|\mathrm{FH}, T > t)$ changes with t (T is the age at disease onset for the targeted individual); 3) the competing risk of death is not negligible.

In this work, we present a modeling of this problem using a Bayesian network mixed with (right-censored) survival outcomes where hazard rates only depend on the genotype of each individual. We explain how belief propagation can be used to obtain posterior distribution of genotypes given the FH, and how to obtain a time-dependent posterior hazard rate for any individual in the pedigree. Finally, we use this posterior hazard rate to compute individual risk, with or without the competing risk of death.

Our method is illustrated using the Claus-Easton model for breast cancer (BC). This model assumes an autosomal dominant genetic risk factor such

*corresponding author, gregory.nuel@math.cnrs.fr

as non-carriers (genotype 00) have a BC hazard rate $\lambda_0(t)$ while carriers (genotypes 01, 10 and 11) have a (much greater) hazard rate $\lambda_1(t)$. Both hazard rates are assumed to be piecewise constant with known values (cuts at 20, 30, ..., 80 years). The competing risk of death is derived from the national French registry.

Keywords: piecewise constant hazard, Bayesian network, belief propagation, Hardy-Weinberg, Mendelian transmission.

1 Introduction

Complex diseases with variable age at onset typically have many interacting factors such as the age, lifestyle, environmental factors, treatments, genetic inherited components. The genetic component is generally composed of one or several genes including major genes for which a deleterious mutation rises significantly the risk of the disease and/or minor genes which participation in the disease is moderate by itself.

The mode of inheritance can be monogenic if a mutation in a single gene is transmitted or polygenic if mutations in several genes are transmitted. As an example of a major gene in a complex disease, the BRCA1 gene is well known to be strongly correlated with ovarian and breast cancer since the 90s (Hall et al., 1990; Claus et al., 1994). Carriers of a deleterious mutation in BRCA1 gene have a much higher risk to be affected with relative risks ranging from 20 to 80 but deleterious mutations in BRCA1 gene only explain 5 to 10 % of the disease (Mehrgou and Akouchekian, 2016) as many other implicated known or unknown genes exist along with sporadic cases (cases with no inherited component).

In other rare genetic diseases such as the Transthyretin-related Hereditary Amyloidosis (THA), no sporadic cases are found and therefore the incidence is equal to zero among non-carriers and all affected individuals are necessarily carriers of a deleterious mutation (Plante-Bordeneuve et al., 2003; Alarcon et al., 2009).

The family history (FH) of such diseases is often the first tool for clinicians to detect a family of carriers of a deleterious mutation as any unusual accumulation of cases in relatives leads to suspect a deleterious allele in the family. With the appropriate model and computation, the FH can be used to better target the most appropriate individuals for a genetic testing and / or to identify high-risk individuals who require special attention (monitoring and/or treatments).

The first challenge to compute such a model comes from the fact that genotypes are mostly (if not totally) unobserved and that posterior carrier probability computations must sum over a large number of familial founders' genotypes configurations. Once such computations are carried out, deriving posterior individual

disease risk is also a challenging task since the posterior carrier distribution changes over time and must be accounted for. Finally, for diseases with possibly late age at onset (*e.g.* cancer), the competing risk of death is not negligible and must be accounted for.

A competing risk situation occurs when an event (called a competing event) precludes the occurrence of the event of interest. This is typically the case for late-onset diseases as the risk of death is not negligible for advanced age. Ignoring the risk of death would amount to assume that death cannot happen and would therefore lead to overestimate the cumulative incidence (the probability of having the disease before any time point). Famous examples of such situations include dementia where the patients are of a particularly advanced age and have a high risk of dying as in Jacqmin-Gadda et al. (2014) or Wanneveich et al. (2016), or studies on geriatric patients (see for instance Berry et al., 2010).

Classical familial risk models such as Claus-Easton (Claus et al., 1991; Easton et al., 1993), BOADICEA (Antoniou et al., 2004), or the BayesMendel models (BRCAPRO, MMRpro, PancPRO and MelaPRO, see Chen et al., 2006) do not take into account the competing event of death. As a result, it is likely that individual predictions will tend to be overestimated from these models (De Pauw, 2012). The main result of the present work is that we show how to derive individual risk predictions from the family history while taking into account the competing risk of death, which is a new contribution to the best of our knowledge.

Another interesting point is that, unlike most similar publications, we here provide all the necessary details to integrate the likelihood over the unobserved genotypes and to compute posterior genotype distributions using Bayesian network and sum-product algorithms. One should not that these models and algorithms clearly are often used in the context of genetics (see Lauritzen, 1996; O'Connell and Weeks, 1998; Fishelson and Geiger, 2002; Lauritzen and Sheehan, 2003; Palin et al., 2011, for a few examples), but rarely fully detailed (see Chen et al., 2006, for example).

It should also be noted that the genetics community usually prefers to rely on simple *peeling* algorithms rather than Bayesian network for pedigree computations but the two concepts are in fact totally equivalent, and the sum-product algorithm presented in this paper can indeed be seen as a simple Bayesian network based reformulation of the most general peeling-based algorithm developed so far (Totir et al., 2009).

The paper is organized as follows: firstly, in Section 2.1 we introduce a formal generic Bayesian network model adaptable to any genetic disease with variable age at onset. Secondly, in Section 2.2, we provide in this context all the necessary details to carry belief propagation on this model, and express the marginal posterior carrier distribution using Bayesian network's potentials. Thirdly, in Section 2.3, we

give closed-form formulas for the posterior individual disease risk, and introduce a simple numerical algorithm allowing to take into account the competing risk of death. Finally, in Section 3, all the methods are illustrated with the Claus-Easton model for breast cancer using the disease model and the parameters of Claus et al. (1991); Easton et al. (1993). In particular, individual predictions derived by taking into account the competing risk of death or ignoring it are compared, which emphasizes the importance of properly taking into account competing risk of death in such models.

2 Materials and Methods

In this section, we first introduce our model (Section 2.1) as a Bayesian network. We next explain how to perform belief propagation in order to obtain posterior carrier distributions (Section 2.2). Finally, we provide all the details needed to derive disease risks predictions from these posterior distributions, including taking into account the competitive risk of death (Section 2.3).

2.1 The Bayesian Network

We consider a total of n (related) individuals. With $\mathcal{I} = \{1, \ldots, n\}$, we denote by $\mathcal{F} \subset \mathcal{I}$ the subset of the founders (i.e. individuals without ancestors in the pedigree) and we denote by $\mathcal{I} \setminus \mathcal{F}$ the set of non-founders (i.e. with ancestors in the pedigree). Let $\boldsymbol{X} = (X_1, \ldots, X_n) \in \{00, 01, 10, 11\}^n$ be the genotypic distribution[1] of the whole family, where X_i denotes the genotype of Individual i. Let $\boldsymbol{T} = (T_1, \ldots, T_n) \in \mathbb{R}^n$ be the time vector representing the age at diagnosis of all individuals. The joint distribution of $(\boldsymbol{X}, \boldsymbol{T})$ is given by:

$$\mathbb{P}(\boldsymbol{X}, \boldsymbol{T}) = \underbrace{\prod_{i \in \mathcal{F}} \mathbb{P}(X_i) \prod_{i \in \mathcal{I} \setminus \mathcal{F}} \mathbb{P}\left(X_i | X_{\mathrm{pat}_i}, X_{\mathrm{mat}_i}\right)}_{\text{genetic part}} \times \underbrace{\prod_{i \in \mathcal{I}} \mathbb{P}(T_i | X_i)}_{\text{survival part}} \tag{1}$$

which corresponds to the definition of a Bayesian Network (BN). See Koller and Friedman (2009) for more details. The genetic part of Eq. (1) only relies on the "classical" Mendelian assumption that the distribution of a non-founder genotype only depends on the parental genotypes. The survival part makes the strong assumption that all T_i are conditionally independent given X_i. This assumption is clearly not true when considering any other familial effect on the disease (*e.g.* polygenic effect, environmental exposure, etc.) which is often taken into account

[1] For the sake of simplicity, we consider here a simple bi-allelic gene but multi-allelic genes can obviously be easily considered.

using a familial random effect (often called *frailty* in the survival context). Such familial random effect is for example assumed to account for a polygenic effect in the BOADICEA model (Antoniou et al., 2002, 2004). Note that for the sake of simplicity, the symbol "$\mathbb{P}$" corresponds through the whole paper either to a probability measure or to a density.

The extension of the present model to frailty models such as BOADICEA is clearly possible and, in many ways, quite straightforward. However, for the sake of simplicity, we focus here on a simpler model and will briefly discuss the extension in the conclusion section. However, even with the strong assumption that T_i only depends on X_i, since (the basically unobserved) $\boldsymbol{X}$ has a strong correlation structure within the pedigree, so does $\boldsymbol{T}$.

We can see on Fig. 1 an example of a moderate size (hypothetical) family with a severe history of breast and ovarian cancer. This family has a total of $n = 12$ individuals with $\mathcal{F} = \{1, 2, 3, 4\}$ and $\mathcal{I} \setminus \mathcal{F} = \{5, 6, 7, 8, 9, 10, 11, 12\}$. There is no inbreeding (mating between individuals with a common ancestor) in this family but a mating loop (two families joined more than once by mating) due to the two brothers of the first nuclear family having children with two sisters of the second nuclear family. Such looped pedigree can be tricky to represent and this explains why Individual 7 appears twice (with an identity link) in Fig. 1.

One should note that loops in pedigree are not the same as cycles in the Bayesian networks framework in the sense that the underlying conditional dependence structure of the model remains a proper directed acyclic graph even in the presence of pedigree with loops.

Genetic Part. For the genetic part, we assume that founders' genotypes are distributed according to the Hardy-Weinberg distribution with disease allele frequency f. It means that for any founder $i \in \mathcal{F}$ we have $\mathbb{P}(X_i = 00) = (1-f)^2$, $\mathbb{P}(X_i = 01) = \mathbb{P}(X_i = 10) = f(1-f)$, and $\mathbb{P}(X_i = 11) = f^2$. This assumption is extremely frequent in family genetics and usually reasonable since it corresponds to the stationary distribution we observe in a population under mild assumptions. However, one should note that other distributions can easily be considered if necessary (*e.g.* genotype 11 forbidden because it is lethal). For the non-founder we simply assume a Mendelian transmission of the alleles, but unbalanced transmission patterns can also be considered.

The genetic part of the model can also be easily extended to account for various constraints. For example, the presence of monozygous twins, say individuals i and j, only requires one to add an identity variable between the two genotypes: $I_{i,j} \in \{0, 1\}$ such as $\mathbb{P}(I_{i,j}|X_i, X_j) = \mathbf{1}\{X_i = X_j\}$. Genetic tests (including error or not) can also be incorporated as additional variables G_i such as $\mathbb{P}(G_i|X_i)$ corresponding to the test specificity and sensibility. Finally, assuming lethal genotypes (*e.g.*

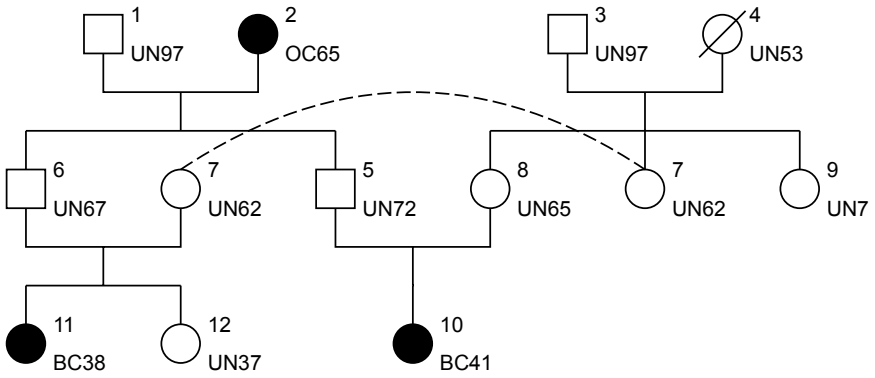


Figure 1: A hypothetical family with a severe FH of cancer. Squares correspond to males, circles to females, and affected individual are filled in black. Individual id on the top-right of the nodes, personal history of cancer (UN=UNaffected; BC=Breast Cancer; OC=Ovarian Cancer) on the bottom-right. The dashed line represents an identity link used to represent the mating loop (due to the mating between individuals 5/8 and 6/7) between brothers 5 and 6, and sisters 7 and 8.

genotype 11) is done straightforwardly by setting to 0 the probability of carrying such genotype. This is equivalent to working conditionally on $\{X_i \neq 11 \text{ for all } i\}$ which obviously alter all genotype distributions, including Hardy-Weinberg for founders.

Survival Part. We place ourselves in the classical survival framework, denoting by $\lambda(t)$ the (time dependent) hazard function, by $S(t)$ the survival function defined as $S(t) = \exp(-\Lambda(t))$ where $\Lambda(t) = \int_0^t \lambda(u)du$ is the cumulative hazard.

We assume an autosomal dominant model where non-carriers have a disease incidence $\lambda_0(t)$ and carriers have a disease incidence $\lambda_1(t)$. This simple assumption results in the following expression of the survival part of the model:

$$\begin{cases} \mathbb{P}(T_i > t | X_i = 00) = S_0(t) \quad \text{and} \quad \mathbb{P}(T_i = t | X_i = 00) = S_0(t)\lambda_0(t) \\ \mathbb{P}(T_i > t | X_i \neq 00) = S_1(t) \quad \text{and} \quad \mathbb{P}(T_i = t | X_i \neq 00) = S_1(t)\lambda_1(t) \end{cases} . \tag{2}$$

As explained above, the symbol "$\mathbb{P}$" corresponds to a (conditional) probability measure for the event $\{T_i > t\}$ and to a density for the punctual event $\{T_i = t\}$.

For example, in the context of the THA, non-carriers cannot be affected ($\lambda_0(t) \equiv 0$) and only carriers have an age-dependent incidence. In the context of breast cancer, $\lambda_0(t)$ might be the incidence for non BRCA carriers and $\lambda_1(t)$ the incidence for BRCA carriers (BRCA1 or BRCA2).

Of course the simple model suggested in Eq. (2) can easily be extended to account for other genetic models (*e.g.* recessive, additive, gonosomal (*i.e.* non-

autosomal), with parent-of-origin effect, etc.) as well as for any known covariates (*e.g.* BMI, smoking, other diseases, etc.) using a classical proportional hazard model.

Hazard rates $\lambda_0(t)$ and $\lambda_1(t)$ are typically described by the literature as piecewise constant hazards (PCHs), but our model allows for any parametric or non-parametric shape as long as hazard rates are provided (*e.g.* hazard rates of Weibull distributions, Gaussian survival, etc.).

2.2 Carrier Risk

For all individual i let us denote by PH_i his/her personal history of the disease. In the case where Individual i was diagnosed with the disease at age t_i we have $\mathrm{PH}_i = \{T_i = t_i\}$. If Individual i was unaffected at age t_i (age at the last follow-up), the variable T_i is right-censored and we have $\mathrm{PH}_i = \{T_i > t_i\}$. From now on, we denote by FH the family history of the disease. This includes the personal history of all individuals and all possible additional constraints or informations (*e.g.* monozygous twins, genetic tests, lethal alleles, etc.). Formally, we can define $\mathrm{FH} = \cup_i(\mathrm{PH}_i \cup \{X_i \in \mathcal{X}_i\})$ where $\mathcal{X}_i \subset \{00, 01, 10, 11\}$ is the subset of allowed values for X_i (*e.g.* $\mathcal{X}_i = \{00, 01, 10\}$ if we know that the genotype 11 is lethal, $\mathcal{X}_i = \{00\}$ if we know that a particular individual is a non-carrier, etc.). Even with genetic testing, it is essential to understand that $\boldsymbol{X}$ is, at best, partially observed. Indeed, even with a (hypothetical and unrealistic) 100% specificity/sensitivity test, a positive heterozygous carrier status cannot distinguish between genotypes 01 and 10. Moreover, genetic tests are in general only available for a few individuals in the whole pedigree. Accounting for the unobserved genotypes is therefore of utmost importance.

Following the classical BN notations, we write the so-called *evidence* $\mathbb{P}(\mathrm{FH})$ as the simple following sum-product of *potentials*:

$$\mathbb{P}(\mathrm{FH}) = \sum_{X_1} \dots \sum_{X_n} \prod_{i=1}^{n} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \tag{3}$$

where the potentials are defined by:

$$K_i\left(X_i | X_{\mathrm{pa}_i}\right) = \mathbb{P}(\mathrm{PH}_i | X_i) \times \begin{cases} \mathbb{P}\left(X_i | X_{\mathrm{pat}_i}, X_{\mathrm{mat}_i}\right) & \text{if } i \in \mathcal{I} \setminus \mathcal{F} \\ \mathbb{P}\left(X_i\right) & \text{if } i \in \mathcal{F} \end{cases} \tag{4}$$

where $\mathbb{P}(\mathrm{PH}_i | X_i)$ is either $\mathbb{P}(T_i = t_i | X_i)$ or $\mathbb{P}(T_i > t_i | X_i)$ and can be obtained through Eq. (2). Note that $\mathrm{pa}_i \subset \mathcal{I}$ denote the parental set of Individual i (empty for founders), and that $X_{\mathcal{J}} = (X_j)_{j \in \mathcal{J}}$ for any $\mathcal{J} \subset \mathcal{I}$. As explained above, any additional information or constraint might and should be added directly into the potentials.

Since $\boldsymbol{X}$ has 4^n possible configurations in the worst case, it is clearly impossible to simply enumerate these configurations even for moderate size pedigrees (e.g., for $n = 10$ or $n = 20$). We therefore need a more efficient algorithm to compute Eq. (3). An efficient solution is provided by the Elston-Stewart algorithm (Elston et al., 1992) in the particular (and frequent) case where the pedigree has no loop. The basic idea is to eliminate variables from the sum-product (*peeling* in the Elston-Stewart literature) from the last generations up to the oldest common ancestor. The resulting complexity $\mathcal{O}(n \times 4^3)$ clearly allows one to deal with arbitrary pedigree size as long as there is no loop.

Unfortunately, loops (inbreeding or mating) are not totally uncommon in pedigrees and therefore have to be accounted for. A simple extension of the Elston-Stewart algorithm consists in using loop breakers: working conditionally to a few number of key genotypes that can be considered as duplicated individuals with known genotypes in a pedigree with no loop. For example, in Fig. 1, Individual 7 is a possible loop breaker. By performing a classical Elston-Stewart algorithm for each genotypic configuration of the loop breakers, $\mathbb{P}(\text{FH})$ can be computed with complexity $\mathcal{O}(n \times 4^{\ell+3})$ where ℓ is the number of loop breakers.

In the context of Bayesian networks, computing $\mathbb{P}(\text{FH})$ (and, in fact, the whole $\mathbb{P}(\boldsymbol{X}, \text{FH})$ distribution) is typically done through *belief propagation* (BP)[2] with a $\mathcal{O}(n \times 4^k)$ complexity where k is the *tree-width* of the graphical model (see Koller and Friedman, 2009, for more details). For a pedigree with no loop, $k = 3$ and the BP complexity is strictly the same than Elston-Stewart, but for more complex pedigrees, k usually increases much slower than $\ell + 3$ and, as a result, BP is often dramatically faster than Elston-Stewart with loop breakers.

In order to achieve this, BP basically eliminates variables from the sum-product of Eq. (3) in a suitable order. In that sense, it is very similar to the notion of *cutset* long used to compute likelihoods in complex pedigrees (see Lange et al., 2013, for a recent reference on the MENDEL package). But BP has the noticeable advantage to allow obtaining the full posterior distribution $\mathbb{P}(\boldsymbol{X}|\text{FH})$ for the same algorithmic complexity while likelihood-based approaches need to repeat many cutset eliminations to achieve the same results. As a consequence, it should not be surprising to see that, in parallel with the classical genetic literature (Elston et al., 1992; Kruglyak et al., 1996; Lange et al., 2013) many authors have been using BP and BN to deal with genetic models (Lauritzen, 1996; O'Connell and Weeks, 1998; Fishelson and Geiger, 2002; Lauritzen and Sheehan, 2003; Palin et al., 2011).

Let us finally point out that the genetics community has put considerable efforts in developing Elston-Stewart algorithms for any Bayesian network counterpart, claiming that *peeling-based* algorithms are more natural for geneticists than junction-tree based ones. Note however that the most general version of these

[2]Also called *sum-product* algorithm.

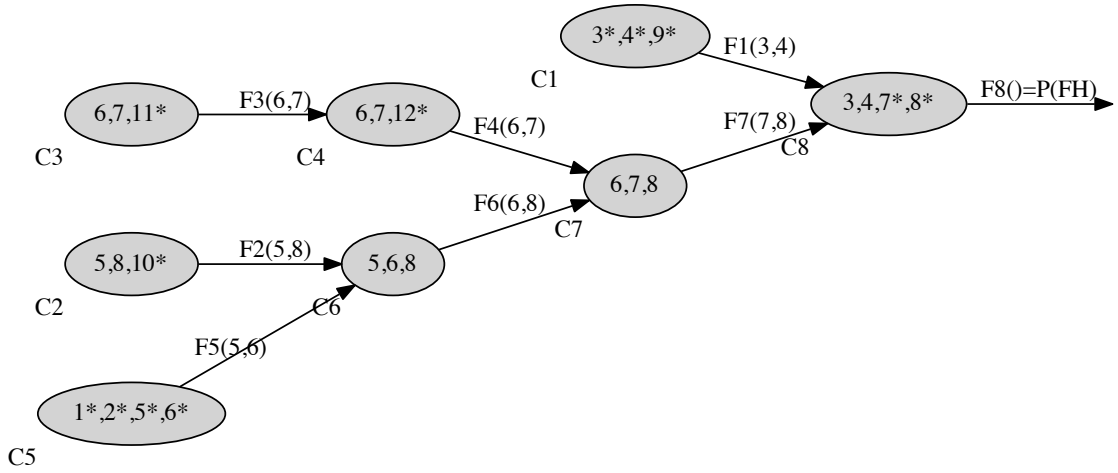


Figure 2: Junction tree of our hypothetical family with the following elimination order: $X_9, X_{10}, X_{11}, X_{12}, X_{1,2}, X_5, X_6, X_{3,4,7,8}$.

peeling algorithms (Totir et al., 2009) is in fact *exactly* equivalent to the classical junction-tree based forward/backward algorithm presented below.

For completeness, we will now briefly recall all the minimal necessary results to implement BP in the context of our model. We nevertheless encourage the interested reader to refer to more classical references like Lauritzen and Sheehan (2003) or Koller and Friedman (2009) for more details.

Variable Elimination and Junction Tree. As an example, we consider the pedigree of Fig. 1 and want to compute $\mathbb{P}(\mathrm{FH})$ by successive variable elimination. We use the following elimination order: X_9, X_{10}, X_{11}, X_{12}, $X_{1,2}$, X_5, X_6, and $X_{3,4,7,8}$. Here follow the quantities obtained in the process:

$$F_1(X_{3,4}) = \sum_{X_9} K_3(X_3)K_4(X_4)K_9(X_{3,4,9});\ F_2(X_{5,8}) = \sum_{X_{10}} K_{10}(X_{5,8,10});$$

$$F_3(X_{6,7}) = \sum_{X_{11}} K_{11}(X_{6,7,11});\ F_4(X_{6,7}) = \sum_{X_{12}} F_3(X_{6,7})K_{12}(X_{6,7,12});$$

$$F_5(X_{5,6}) = \sum_{X_{1,2}} K_1(X_1)K_2(X_2)K_5(X_{1,2,5})K_6(X_{1,2,6});\ F_6(X_{6,8}) = \sum_{X_5} F_2(X_{5,8})F_5(X_{5,6});$$

$$F_7(X_{7,8}) = \sum_{X_6} F_4(X_{6,7})F_6(X_{6,8});\ \mathbb{P}(\mathrm{FH}) = \sum_{X_{3,4,7,8}} F_1(X_{3,4})F_7(X_{7,8})K_7(X_{3,4,7})K_8(X_{3,4,8}).$$

We therefore can obtain $\mathbb{P}(\mathrm{FH})$ by considering only $6 \times 4^3 + 2 \times 4^4 = 896$ configurations over the $4^{12} \simeq 16.8 \times 10^6$ total number of $\boldsymbol{X}$ configurations. Note that a memory bounded version of the variable elimination exists, see Darwiche (2001) for more details.

Fig. 2 is a graphical representation of this particular sequence of elimination and is also a *junction tree* defined as a set of K *cliques* $C_1, \ldots, C_K$ with $C_j \subset \{X_1, \ldots, X_n\}$ with the following properties:

i) tree: each clique j is connected to a a subsequent *clique* $\mathrm{to}_j \in \{j+1, \ldots, K\}$ ($\mathrm{to}_K = \text{root}$ by convention). We also define $\mathrm{from}_k = \{j, \mathrm{to}_j = k\}$ ($\mathrm{from}_1 = \emptyset$) and $S_j = C_j \cap C_{\mathrm{to}_j}$ (with the convention that $S_K = \emptyset$).

ii) covering: for all $i \in \{1, \ldots, n\}$ it exists a j such as $\{X_i, X_{\mathrm{pa}_i}\} \subset C_j$. We then define $\mathrm{of}_i = \min\{j, (X_i, X_{\mathrm{pa}_i}) \subset C_j\}$ and $C_j^* = \{X_i \in C_j, \mathrm{of}_i = j\}$.

iii) running intersection: for all $i \in \{1, \ldots, n\}$ the subgraph formed by $\{C_j, X_i \in C_j\}$ (and the from/to relationships) is a tree.

In the graph theory, junction trees are used as an auxiliary structure for many applications (*e.g.* graph coloring). The proof that any elimination sequence gives a junction tree can be found in Koller and Friedman (2009). The *tree-width* of an elimination sequence / junction tree is defined as the size of its largest clique. Finding the elimination sequence with the smallest tree-width is NP-hard in general, but many heuristics are available (Koller and Friedman, 2009). The elimination order of Fig. 2 has been obtained using the well-known minimum fill-in heuristic.

Belief Propagation. We assume that a suitable elimination order / junction tree has been obtained. For all $j \in \{1, \ldots, K\}$ we hence define the potential of clique C_j as $\Phi_j(C_j) = \prod_{X_i \in C_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right)$ and we have the following result:

Theorem 1. *(posterior distribution) For all $i \in \{1, \ldots, n\}$, let $k = \mathrm{of}_i$ and we have:*

$$\mathbb{P}(X_i, \mathrm{FH}) = \sum_{C_k \setminus \{X_i\}} \left\{ \prod_{j \in \mathrm{from}_k} F_j(S_j) \times \Phi_k(C_k) \times B_k(S_k) \right\}$$

where the forward *quantities are defined for $k = 1, \ldots, K$ by:*

$$F_k(S_k) = \sum_{C_k \setminus S_k} \left\{ \prod_{j \in \mathrm{from}_k} F_j(S_j) \times \Phi_k(C_k) \right\}$$

and the backward *quantities are defined by $B_K(S_K = \emptyset) = 1$ (convention) and for $k = K, \ldots, 2$, for all $i \in \mathrm{from}_k$:*

$$B_i(S_i) = \sum_{C_k \setminus S_i} \left\{ \prod_{j \in \mathrm{from}_k, j \neq i} F_j(S_j) \times \Phi_k(C_k) \times B_k(S_k) \right\}.$$

Proof. See Appendix A. □

Using Theorem 1, it is therefore possible to obtain $\mathbb{P}(\text{FH})$ and *all* $\mathbb{P}(X_i|\text{FH})$ by just recursively computing once all forward and backward quantities.

2.3 Disease Risk

While the previous section covered the computation of the posterior probability $\mathbb{P}(X_i|\text{FH})$ for all individuals in the pedigree, we now focus in this section on computing individual posterior disease risks, with or without the competing risk of death.

Risk without competing events. We consider an individual i with a posterior carrier probability π at age τ, that is $\pi = \mathbb{P}(X_i \neq 00|\text{FH}, T_i > \tau)$. Conditionally to the family history, we denote the survival and hazard functions respectively by S and λ such that, for $t \geq \tau$, $S(t) = \mathbb{P}(T_i > t|\text{FH}, T_i > \tau)$ and $S(t) = \exp(-\int_\tau^t \lambda(u)du)$. We have the following result.

Theorem 2. *For any $t \geq \tau$, we have:*

$$S(t) = \pi \frac{S_1(t)}{S_1(\tau)} + (1-\pi)\frac{S_0(t)}{S_0(\tau)}$$

$$\mathbb{P}(X_i \neq 00|\text{FH}, T_i > t) = \frac{1}{S(t)}\pi\frac{S_1(t)}{S_1(\tau)}$$

$$\lambda(t) = \frac{1}{S(t)}\left[\pi\frac{S_1(t)}{S_1(\tau)}\lambda_1(t) + (1-\pi)\frac{S_0(t)}{S_0(\tau)}\lambda_0(t)\right] \tag{5}$$

Proof. See Appendix B. □

Risk with death as a competing event. As explained in the introduction, death precludes the occurence of the disease. This needs to be taken into account by defining the hazard rate of the disease conditionally to the fact that both disease and death have not occurred yet. From a statistical point of view, such a situation can be seen as a competing risk situation or as an illness-death model; see Andersen et al. (1993) or Andersen and Keiding (2012) for a presentation of such models. We define T^* as the minimum between age at disease onset and age at death and we keep the notation T to denote the age at disease onset. Given an individual i with a family history FH, its hazard rate for the disease is defined as

$$\lambda_\alpha(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t | T_i^* \geq t, \text{FH})}{\Delta t}$$

We denote by λ_β and S_β the hazard and survival functions of T_i^* (conditionally to the family history) and we assume that λ_α and λ_β are piecewise constants with common cuts $\tau = c_0 < c_1 < \ldots < c_N$ (that is $\lambda_\alpha(t) = \alpha_j$ and $\lambda_\beta(t) = \beta_j$ for $t \in]c_{j-1}, c_j]$).

Lemma 3. *For $j = 1, \ldots, N$, $t \in]c_{j-1}, c_j]$, we have*

$$\mathbb{P}(T_i \leq t | T_i > c_{j-1}, \mathrm{FH}) = \int_{c_{j-1}}^{t} \lambda_\alpha(u) S_\beta(u) du = \frac{\alpha_j}{\beta_j} \left[S_\beta(c_{j-1}) - S_\beta(t) \right]$$

Proof. See Appendix B. □

Practical computations. We assume that one individual has a carrier probability π at age τ (his age without the disease in the FH). We denote by λ_{death} his/her hazard of death. Then the posterior disease risk with the competing risk of death can be computed through the following steps:

1) choose a fine enough discretization $\tau = c_0 < c_1 < \ldots < c_N = t_{\max}$ (ex: all $c_j - c_{j-1} = 0.1$ year);

2) compute $\alpha_j = \lambda_\alpha(c_j)$ using Eq. (5);

3) compute $\beta_j = \alpha_j + \lambda_{\mathrm{death}}(c_j)$;

4) then the marginal posterior probability of being diagnosed with the disease before age c_k, in the presence of death as a competing risk, is given for $k = 1, \ldots, N$ by:

$$\mathbb{P}(T_i \leqslant c_k | \mathrm{FH}) = \sum_{j=1}^{k} \frac{\alpha_j}{\beta_j} \left[S_\beta(c_{j-1}) - S_\beta(c_j) \right].$$

3 Results and Discussion

3.1 The Claus-Easton Model

In order to illustrate our method, we will use the model of illness and the parameters of the Claus-Easton model developed from the Cancer and Steroid Hormone Study in the 90s (Claus et al., 1991; Easton et al., 1993).

The Claus-Easton model is a classical genetic model composed of a genotypic part and a phenotypic part with only the family history (FH) as covariate. It assumes an autosomal dominant mode of inheritance, and a piecewise constant hazard rate by steps of 10 years. The penetrance ($F(t) = 1 - S(t)$) and the

	$20-30$	$30-40$	$40-50$	$50-60$	$60-70$	$70-80$	>80
non carriers	2.00	26.04	112.94	139.94	235.17	232.16	232.03
carriers	168.35	1391.49	3153.21	3222.22	3281.25	3289.86	3286.43
relative risk	84.17	53.44	27.92	23.03	13.95	14.17	14.16

Table 1: Annual incidence (for 100,000) of breast cancer (BC) for carriers/non-carriers and relative risks by age (in years) in the Claus-Easton model.

$20-30$	$30-40$	$40-50$	$50-60$	$60-70$	$70-80$	$80-85$
23.85375	46.86641	130.5396	308.9539	599.914	1493.6	3845.406
$85-90$	$90-95$	$95-99$	$99-100$	$100-101$	$101-102$	$102-103$
8114.203	16400.99	27912.22	35644	38696.22	43033.07	45647.85

Table 2: Annual female death incidence (for 100,000) by age (in years) in the metropolitan French population between 2012 and 2014 (INED, 2017).

density ($f(t) = \lambda(t)S(t)$) are given in Table 2 from Easton et al. (1993) for both carriers and non-carriers at ages $25, 35, \ldots, 85$. The hazard rates can therefore be derived from these data using the formula $\lambda(t) = f(t)/(1 - F(t))$. The results of these computations are given in Table 1. The frequency of the mutated allele has been estimated at $f = 0.0033$ (Claus et al., 1991). The death incidences needed in the competing risk section are given in Table 2.

Figure 3 presents the incidence and survival for BC (carriers and non-carriers) as well as death. We can notice that the breast cancer incidences in carriers are always much higher than in non-carriers at any age and the relative risk between carriers and non-carriers is especially large (RR > 50) before age 40 (see Table 1) but then decreases with aging. We notice that the death incidence stays above the BC incidence for non-carriers at all ages and exceeds even the BC incidence for carriers from age 80. This shows the importance of taking it into consideration especially over a certain age.

3.2 Carrier Risk

In this section we will use the belief propagation in Bayesian networks to obtain the posterior distribution of individual genotypes given the FH. We get the posterior probabilities of each genotype (non-carrier, heterozygous carrier with a paternal mutated allele, heterozygous carrier with a maternal mutated allele and homozygous carrier).

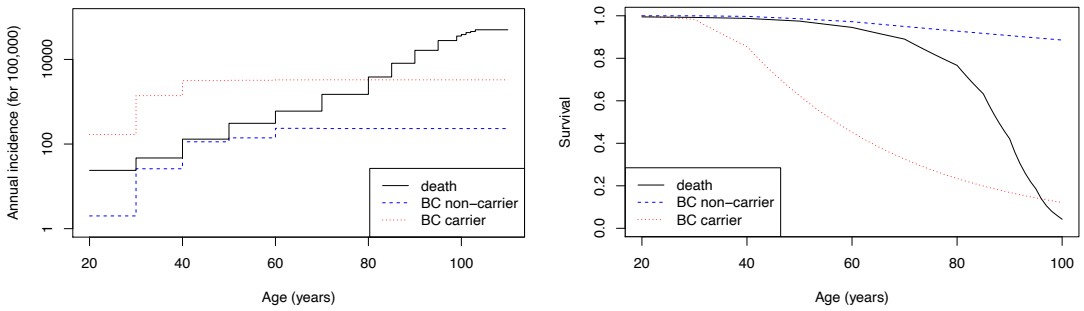


Figure 3: Left-panel: annual (female) death incidence and annual non-carrier/carrier breast cancer incidence. Right-panel: death survival and percentage of non-carrier/carrier individuals without diagnosed breast cancer.

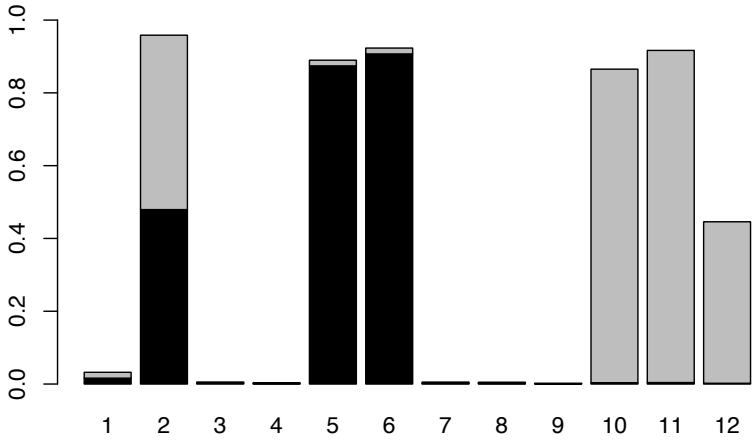


Figure 4: Posterior probabilities for the carrier genotypes of each individual (Individual 1 to 12) in our hypothetical family (Figure 1). The posterior probability of being a paternal carrier $\mathbb{P}(X = 10|\text{FH})$ (resp. maternal carrier $\mathbb{P}(X = 01|\text{FH})$) is colored in black (resp. in grey). The deleterious allele being very rare in the general population ($f = 0.33\%$), the probability of the monozygous carrier genotype is almost zero for each individual and it is therefore not represented here.

Figure 4 represents the marginal posterior probability $\mathbb{P}(X_i = x|\text{FH})$ for all individuals i and for $x = 10$ (paternal carrier) and $x = 01$ (maternal carrier). Note that the posterior probability of the monozygous carrier genotype ($x = 11$) being almost zero for each individual, it is not shown here. The posterior probability of the non-carrier genotype can be easily deduced.

We can notice that the probabilities of being a non-carrier for 1, 3, 4, 7, 8 and 9 are all by far the highest despite the severe phenotype of relatives (granddaughter, niece or daughter). This result is consistent with the personal history of Individual 2 (ovarian cancer at age 51) which points her out as the most likely origin of the mutation in the family. Let us note that since we have no additional information on the ancestors of Individual 2, it is impossible to determine whether her mutation was transmitted by her father or her mother. As a consequence, the posterior carrier probability is equally shared between the paternal and maternal carrier genotypes.

Considering the severe personal history of cancer of Individuals 10 and 11, the most likely situation would be that they both received the mutation of their grandmother through their respective fathers (Individuals 6 and 5 respectively). The posterior probabilities are clearly consistent with this scenario: Individuals 5 and 6 have a probability of $\simeq 90\%$ to be maternal carriers, and Individuals 10 and 11 have similar probabilities to be paternal carriers. Note that Individual 12, being unaffected at age 37 (which is not very informative) basically have 50% chance to have received the mutation from her father.

Figure 5 shows some examples of the variations of the posterior marginal distribution of the genotypes in a same family structure according to different FH. We first notice that with no information (FH1) the posterior probabilities are exactly those of the general population: $\mathbb{P}(X_i \neq 00|\text{FH1}) = 1 - (1 - f)^2 \simeq 0.0066$.

Note that Individual 2 has a severe personal history of cancer (ovarian cancer at age 51) in all other examples. As a consequence, Individual 1, as a male with no personal history of cancer, is mostly totally uninformative therefore not included in the forthcoming analyses.

Individual 4 having no children, she is independent from the rest of the family conditionally to her phenotype and her parent's genotype. With no information about her phenotype in any FH, her probability of being a carrier is therefore almost half her mother's one in each FH (because her father is almost uninformative). If we compare the posterior distribution of the genotype of Individual 3 in FH2, FH3 and FH4, we can notice that the ovarian cancer of her mother which increased her mother's probability of being a carrier raises her probability of being a carrier (FH2). A protective information about her phenotype such as no cancer until age 61 lowers her posterior probability of being a carrier (FH3). On the contrary, the cancer at young age of her daughter which increases her daughter's

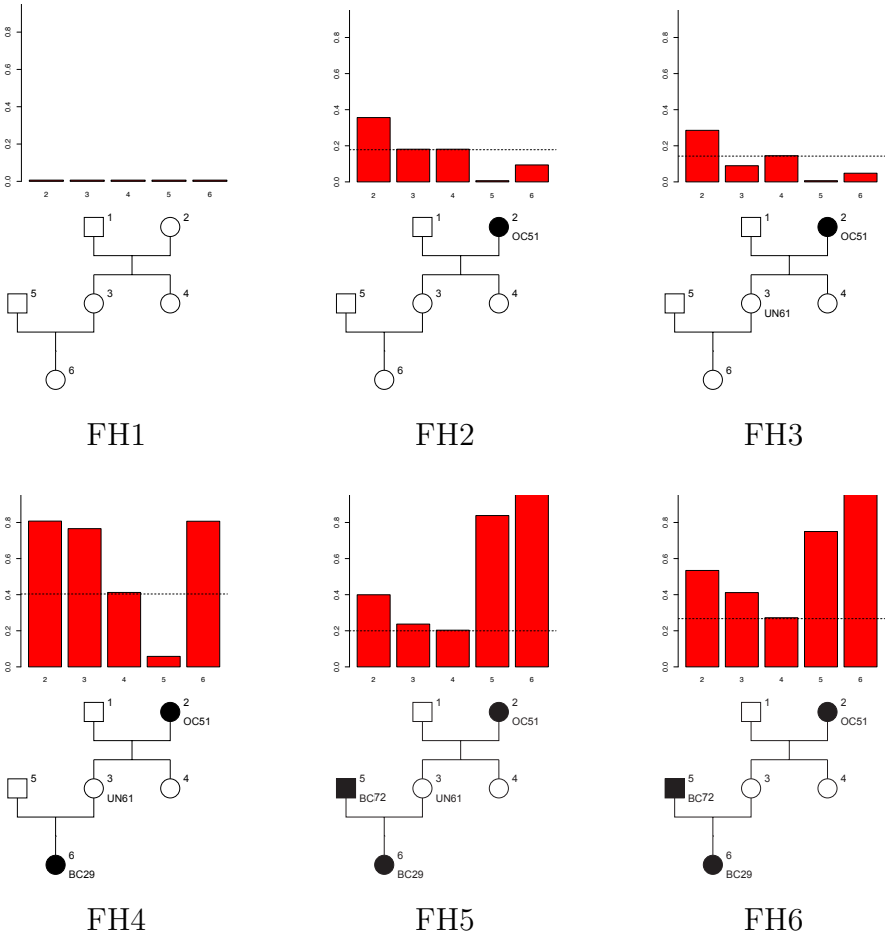


Figure 5: Posterior marginal carrier distribution for a total of 6 FH with increasing degree of severity on the same pedigree structure with 6 individuals. Dashed line represent half the marginal carrier probability of Individual 2.

probability of being a carrier raises her own probability of being a carrier (FH4-6).

We also notice the causal relationships in a whole branch of the family with the transmission between Individuals 2, 3 and 6 of the deleterious allele being highly probable which raises the probability of being a carrier for Individual 3 even in the presence of a protective phenotype (unaffected at age 61) in FH4.

We finally observe the influence of the spouse's genotype when having children (FH5). The higher risk of being a carrier for Individual 5 (because of his cancer at age 72) strongly decreases the carrier probability of his spouse (in comparison with FH4) since the paternal origin of the disease mutation naturally becomes the most likely event. On the other side, the increase of risk for Individual 3 when suppressing her protective phenotype (FH6) also has a consequence on the marginal posterior distribution of her spouse in lowering his probability of being a carrier as his participation in the risk for their daughter is lowered.

To summarize, one's probability of being a carrier mainly depends on: 1) one's probability of having at least one carrier parent, which is correlated to the history of cancer of one's ancestors; 2) one's probability of having transmitted the mutation to one's offspring which is correlated to the history of cancer of one's descendant relatives and one's spouse probability of being a carrier.

Remark: As introduced in the "Disease Risk" section, we know that posterior carrier probabilities should decrease with time for unaffected individuals. For example, if we assume that Individual 4 is unaffected at age 40 in FH6, her probability of being a carrier is 24%. If she stays unaffected up to age 60 (resp. age 80), her probability of being a carrier decreases to 15% (resp. 8.5%).

Table 3 gives a practical illustration of the dependence and conditional independence in a trio grandparent - parent - child. We compare the posterior joint distribution and the product of the posterior marginal distributions of genotypes X_2 and X_6 in FH4 with various information on X_3. We can see that these two quantities are not equal when X_3 is not observed while they are exactly the same when X_3 is fixed. This example demonstrates how X_2 and X_6 are not conditionally independent given FH but they are, conditionally to FH and X_3. Note that when $X_3 = 11$, the mutation is necessarily found in both parents (Individual 1 and 2) as well as in her daughter (Individual 6).

3.3 Cancer Risk

As in Section 2.3 we now consider a female individual i who is unaffected at age τ (*i.e.* $\{T_i > \tau\} \subset \text{FH}$) and denote by $\pi = \mathbb{P}(X_i \neq 00|\text{FH})$ its posterior carrier probability. The purpose of this section is to compute the posterior risk of cancer for this individual (with or without the competing risk of death). As previously explained, these risks only depend on π and τ.

X_2/X_6	NC/NC	NC/C	C/NC	C/C
FH4				
marginal	0.0371306	0.1551811	0.1559446	0.6517438
joint	0.1443102	0.0480015	0.0487650	0.7589233
FH4 and $X_3 = 10$				
marginal	0.0092840	0.7741949	0.0025657	0.2139554
joint	0.0092840	0.7741949	0.0025657	0.2139554
FH4 and $X_3 = 01$				
marginal	0.0000000	0.0000000	0.0118497	0.9881503
joint	0.0000000	0.0000000	0.0118497	0.9881503
FH4 and $X_3 = 11$				
marginal	0.0000000	0.0000000	0.0000000	1.0000000
joint	0.0000000	0.0000000	0.0000000	1.0000000

Table 3: product of the posterior marginal probabilities $\mathbb{P}(X_2|\text{FH})\mathbb{P}(X_6|\text{FH})$ and joint posterior probability $\mathbb{P}(X_2, X_6|\text{FH})$ in the context of known and unknown X_3. NC: non-carrier; C: carrier.

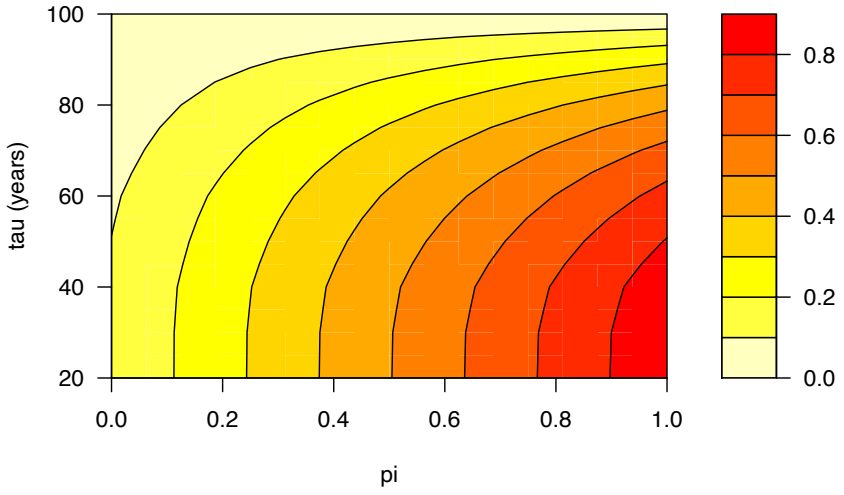


Figure 6: Individual risk of breast cancer without the competing risk of death and for various π and τ

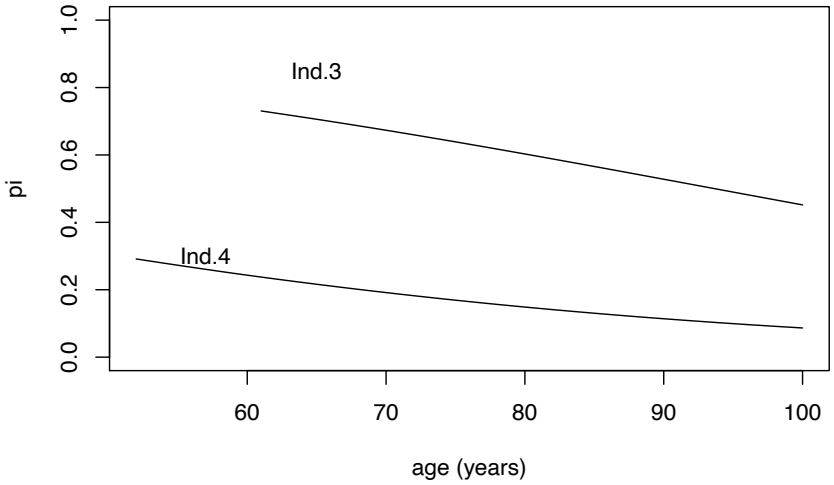


Figure 7: Posterior probabilities of being a carrier according to the time for Individuals 3 and 4 in FH4 assuming Individual 4 is 52 at the time of the censoring.

Figure 6 represents the individual risk of breast cancer up to age 100[3] without the competing risk of death and variant π and τ. We can see that the individual risk of BC rises as π increases and τ decreases. This result is quite intuitive as the younger a patient is, the longer she will be at risk until age 100; the greater her probability of carrying a deleterious allele, the greater her risk to develop a cancer.

As introduced in the previous section the probability of being a carrier for an unaffected individual decreases with time if she stays unaffected. Assuming Individual 4 was 52 in FH4, Figure 7 shows the evolution of the probability of being a carrier for Individual 3 and Individual 4 in FH4. As they stay unaffected we can clearly see the decrease of this probability which has to be taken into account in the computation of the individual risk of breast cancer over time (see Section 2.3).

As explained in Section 2.3, computing risk with the competing risk of death requires a numerical discretization of age by a fixed step Δt. In order to calibrate Δt we used $\Delta t = 0.01$ as a reference, and observed that $\Delta t = 0.1$ is a reasonable balance between accuracy and computational efficiency (data not shown).

Figure 8 represents the individual risk of breast cancer for Individual 7 ($\pi =$

[3]Note that we obtain qualitatively similar results with a lower age limit (*e.g.* age 80), but quantitative results are more illustrative with age 100.

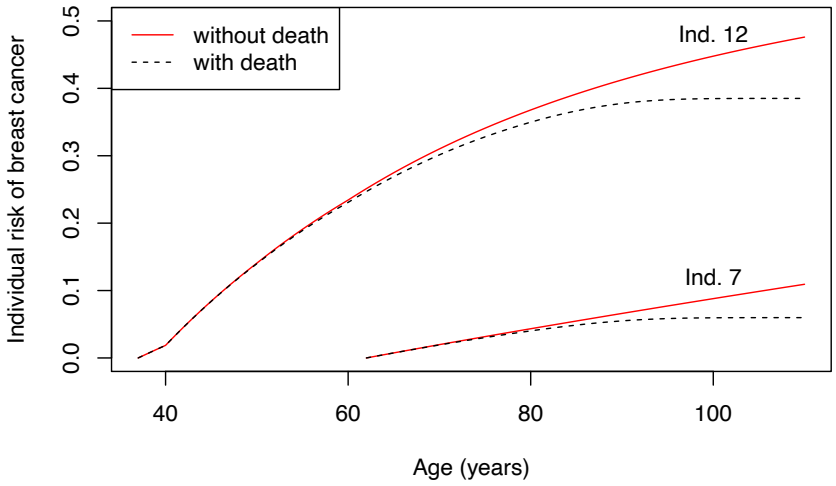


Figure 8: Individual risk of breast cancer with and without the competing risk of death for individual 7 and 12 of our hypothetical family from τ to 100 years with and without the competing risk of death.

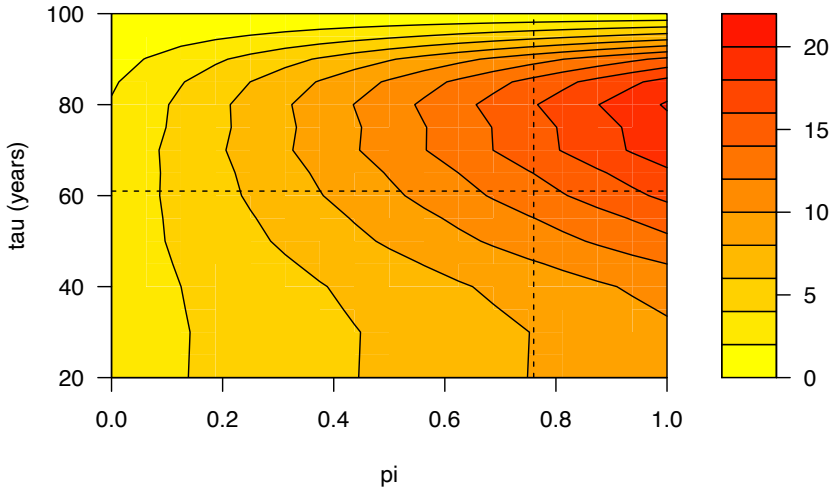


Figure 9: Difference (in percentage) between the individual risk of breast cancer up to 100 years without and with the competing risk of death for various π and τ. Specific values $\pi = 76.59\%$ and $\tau = 61$ are given by the dashed lines.

0.553% and $\tau = 62$ years) and Individual 12 ($\pi = 44.6\%$ and $\tau = 37$ years) in our hypothetical family from τ to 100 years with and without taking into account the competing risk of death. We can see that the difference between the two curves for each individual is increasing with the age. The age from which the difference becomes significant varies with the couple (π, τ). We also observe that the individual risk of breast cancer eventually reaches a plateau which corresponds to the point where the incidence of breast cancer becomes negligible compared to the incidence of death in the elderly.

Quantitatively, the importance of taking into account the competing risk of death is pointed out in the Figure 9 which represents the difference between the individual risk of breast cancer up to the age of 100 years for variant couples (π, τ). For example for Individual 3 in FH4 ($\pi = 76.59\%$, $\tau = 61$, see Figure 5), the error while calculating her individual risk of breast cancer up to the age of 100 years reaches almost 14 %. If it is clear that the competing risk of death can have a limited effect on the global risk of cancer for certain couples (π, τ) its effect is never totally negligible, and since we provide a rigorous way to take it into account we strongly advocate its use in all circumstances.

4 Conclusions

We presented here a general model for genetic disease with variable age at onset. This model, a Bayesian network, combines classical genetic modeling with survival analysis. In order to deal with the (mostly) unobserved genotypes, we first explained in detail how belief propagation can be used to perform likelihood and posterior probability computations. Secondly, we focused on the challenging problem of computing posterior individual disease risks, with or without taking into account the competing risk of death. Finally, we illustrated theses results with the Claus-Easton model for breast and ovarian cancer. The R source codes are available upon request for the interested readers.

For the sake of simplicity, we only considered a bi-allelic locus with standard distribution (autosomal, Hardy-Weinberg, Mendelian allele transmission) but extensions (*e.g.* multi-loci, unbalanced allele transmission, lethal genotypes, etc.) are straightforward. For the survival model, we presented a simple dominant effect without covariates, but again, extensions to any proportional hazard model (*e.g.* recessive, additive, with covariates, etc.) are easy to implement. Incorporating random effects (at the individual and/or familial level) in the model (like in the BOADICEA model, see Antoniou et al., 2002, 2004) is clearly also possible, but slightly more challenging.

Computation of posterior carrier distributions remains almost unchanged except for the random effect support which must be discretized (five values are claimed to be sufficient in the BOADICEA literature) and for the belief propagation which must be performed once for each of the possible value of the random effect. For posterior risks, calculations get slightly more complex since the posterior individual hazard must now be integrated over the (changing over time) posterior joint distribution of the individual genotype and of the random effect. Basically, all computations are slightly more intensive with random effects, but most results of Section 2.3 remain very similar.

One of the important limitations of the present work is the fact that we assume that all model parameters are known. However, it should be noted that likelihood and conditional likelihood might be easy to compute through the belief propagation which means that we basically provide all the necessary means to estimate the model parameters from actual data. In that context, it is nevertheless critical to deal efficiently with ascertainment issues: the fact that the family ending up in the database are usually precisely the one with the most severe disease family history. But standard methods like the PEL (Alarcon et al., 2009), which basically are conditional likelihood computations, are known to deal relatively well with the problem.

In order to take into account the competing risk of death, we used death from all causes, which was obtained from registry data (INED, 2017). However, only

death without cancer precludes the onset of cancer and we are not interested into death from all causes. Since registry data usually do not report the causes of death it is a difficult task to estimate the risk of death without cancer. This has been studied for instance in Wanneveich et al. (2016) through a illness-death model, using registry data and differential equations to model the specific causes of death. Nevertheless, it is very likely that the gain in terms of predictions would be minor as mortality from all causes is likely to be close to mortality without cancer.

Further work includes all the extensions described above (*e.g.* more complex genetic model, genetic tests, familial random effects, etc.) as well as the development of a clinical web application for the Claus-Easton model in close collaboration with the cancer genetics department of the *Institut Curie*. From the methodological point of view, we plan to focus on the computation of more complex posterior distribution like the number of carriers in any subgroup of individuals and/or the familial posterior risk (time before any family member at risk is diagnosed).

Acknowledgments

We would first like to thank both anonymous reviewers for their constructive comments and remarks. This work received the support of both the *Institut de Recherche en Santé Publique* (IRESP) and the *Ligue National Contre le Cancer* (LNCC). Alexandra Lefebvre's internship was funded by the *Institut Curie*. We finally warmly thank Antoine de Pauw (*Institut Curie*) for his continuous friendly support and for suggesting the hypothetical family presented here.

Conflict of Interest

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

A Proofs for the Carrier Risk Section

For all $k \in \{1, \dots, K\}$ we recursively define: $u_k = \{k\} \cup_{j \in \text{from}_k} u_j$, $U_k = \cup_{j \in u_k} C_j$, and $V_k = \cup_{j \notin u_k} C_j$. Then we can compute the so-called *forward* and *backward* quantities over any separator $S_j = C_j \cap C_{\text{to}_j}$:

$$F_j(S_j) = \sum_{U_j \setminus S_j} \prod_{X_i \in U_j^*} K_i\left(X_i | X_{\text{pa}_i}\right) \quad \text{and} \quad B_j(S_j) = \sum_{V_j \setminus S_j} \prod_{X_i \in V_j^*} K_i\left(X_i | X_{\text{pa}_i}\right)$$

where $U_j^* = \{X_i \in U_j, \exists k \in u_j, \text{of}_i = k\}$ and $V_j^* = \{X_i \in V_j, \exists k \notin u_j, \text{of}_i = k\}$.

The key is then to prove that, for all $j \in \{1, \ldots, K\}$ we have:

$$\mathbb{P}(S_j, \mathrm{FH}) = F_j(S_j) B_j(S_j) \tag{6}$$

$$\mathbb{P}(C_k, \mathrm{FH}) = \Phi_k(C_k) \times \prod_{j \in \mathrm{from}_k} F_j(S_j) \times B_k(S_k). \tag{7}$$

For proving Eq. (6), we start by noticing that the JT (Junction Tree) properties (Koller and Friedman, 2009) give: $\{X_1, \ldots, X_n\} \setminus S_j = (U_j \setminus S_j) \uplus (V_j \setminus S_j)$ and $\{X_1, \ldots, X_n\} = U_j^* \uplus V_j^*$ (both being disjoint unions). We therefore have:

$$\begin{aligned}
\mathbb{P}(S_j, \mathrm{FH}) &= \sum_{U_j \setminus S_j} \sum_{V_j \setminus S_j} \prod_{X_i \in U_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \prod_{X_i \in V_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \\
&= \underbrace{\left(\sum_{U_j \setminus S_j} \prod_{X_i \in U_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \right)}_{F_j(S_j)} \times \underbrace{\left(\sum_{V_j \setminus S_j} \prod_{X_i \in V_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \right)}_{B_j(S_j)}
\end{aligned}$$

the factorization between the first and second equation being possible thanks to the fact that $\left(\cup_{X_i \in U_j^*} \{X_i, X_{\mathrm{pa}_i}\}\right) \cap \left(\cup_{X_i \in V_j^*} \{X_i, X_{\mathrm{pa}_i}\}\right) = S_j$ (JT properties again).

The proof is basically the same for Eq. (7) using $\{X_1, \ldots, X_n\} \setminus C_k = \uplus_{j \in \mathrm{from}_k} (U_j \setminus S_j) \uplus (V_k \setminus S_k)$ we get:

$$\begin{aligned}
\mathbb{P}(C_k, \mathrm{FH}) &= \sum_{\{X_1, \ldots, X_n\} \setminus C_k} \prod_{X_i \in \{X_1, \ldots, X_n\}} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \\
&= \Phi_k(C_k) \prod_{j \in \mathrm{from}_k} \sum_{U_j \setminus S_j} \sum_{V_k \setminus S_k} \prod_{X_i \in U_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \prod_{X_i \in V_k^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right) \\
&= \Phi_k(C_k) \prod_{j \in \mathrm{from}_k} \underbrace{\sum_{U_j \setminus S_j} \prod_{X_i \in U_j^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right)}_{F_j(S_j)} \underbrace{\sum_{V_k \setminus S_k} \prod_{X_i \in V_k^*} K_i\left(X_i | X_{\mathrm{pa}_i}\right)}_{B_k(S_k)}.
\end{aligned}$$

The factorisation being possible as $\uplus_{j \in \mathrm{from}_k} (U_j \setminus S_j) \cap (V_k \setminus S_k) = \emptyset$ (running intersection) and $\forall j, \forall k, U_j^* \subseteq U_j$ and $V_k^* \subseteq V_k$.

Finally, the recursive expression of the forward and backward quantities can be easily derived from equations (6) and (7):

$$\begin{aligned}
\mathbb{P}(S_k, \mathrm{FH}) &= \sum_{C_k \setminus S_k} \mathbb{P}(C_k, \mathrm{FH}) \\
F_k(S_k) \cancel{B_k(S_k)} &= \sum_{C_k \setminus S_k} \prod_{j \in \mathrm{from}_k} F_j(S_j) \times \Phi_k(C_k) \times \cancel{B_k(S_k)}
\end{aligned}$$

which gives the forward recursion by simplifying the $B_k(S_k)$ term.

B Proofs for the Disease Risk Section

Proof of Theorem 2. For clarity, we recall that $S_0(t) = \mathbb{P}(T_i > t|X_i = 00)$, $S_1(t) = \mathbb{P}(T_i > t|X_i \neq 00)$, $\pi = \mathbb{P}(X_i \neq 00|\text{FH}, T_i > \tau)$ and $S(t) = \mathbb{P}(T_i > t|\text{FH}, T_i > \tau)$, for $i = 1, \ldots, n$, and that $\{T_i > \tau\} \subset \text{FH}$. Since the T_i are independent conditionally to the X_i, the distribution of T_i conditionally on X_i obviously does not depend on FH (for values of X_i which are not forbidden by FH). This is why FH can be omitted almost everywhere in the following proof as soon as π has been computed.

We have $S(t) = \sum_{X_i} \mathbb{P}(T_i > t, X_i|T_i > \tau, \text{FH})$, where the notation $\sum_{X_i}$ represents the summation over the different possible values of X_i, that is $X_i = 00$ or $X_i \neq 00$. Using Bayes' rule,

$$\begin{aligned}\mathbb{P}(T_i > t, X_i \neq 00|T_i > \tau, \text{FH}) &= \mathbb{P}(T_i > t|X_i \neq 00, T_i > \tau, \text{FH}) \times \mathbb{P}(X_i \neq 00|T_i > \tau, \text{FH}) \\ &= \frac{\mathbb{P}(T_i > t, X_i \neq 00, \text{FH})}{\mathbb{P}(T_i > \tau, X_i \neq 00, \text{FH})} \times \pi \\ &= \frac{\mathbb{P}(T_i > t|X_i \neq 00, \text{FH})}{\mathbb{P}(T_i > \tau|X_i \neq 00, \text{FH})} \times \pi = \frac{S_1(t)}{S_1(\tau)}\pi,\end{aligned}$$

where we used the fact that $\mathbb{P}(T_i > t|X_i \neq 00, \text{FH}) = \mathbb{P}(T_i > t|X_i \neq 00)$. We similarly prove that $\mathbb{P}(T_i > t, X_i = 00|T_i > \tau, \text{FH}) = (1 - \pi)S_0(t)/S_0(\tau)$.

The next result is proved using Bayes' rule:

$$\begin{aligned}\mathbb{P}(X_i \neq 00|\text{FH}, T_i > t) &= \frac{\mathbb{P}(X_i \neq 00, \text{FH}, T_i > t)}{\mathbb{P}(\text{FH}, T_i > t)} \\ &= \frac{\mathbb{P}(T_i > t|X_i \neq 00, T_i > \tau)}{\mathbb{P}(T_i > t|\text{FH}, T_i > \tau)}\mathbb{P}(X_i \neq 00|\text{FH}, T_i > \tau),\end{aligned}$$

where we also used the fact that $\mathbb{P}(T_i > t|X_i \neq 00, \text{FH}, T_i > \tau) = \mathbb{P}(T_i > t|X_i \neq 00, T_i > \tau)$.

We then directly have $\mathbb{P}(T_i > t|X_i \neq 00, T_i > \tau) = S_1(t)/S_1(\tau)$ from Bayes' rule, $\mathbb{P}(X_i \neq 00|\text{FH}, T_i > \tau) = \pi$ and $\mathbb{P}(T_i > t|\text{FH}, T_i > \tau) = S(t)$ which concludes the proof.

Finally, in order to prove Equation (5), we recall that

$$\begin{aligned}\lambda(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t|T_i \geq t, \text{FH})}{\Delta t} \\ \lambda_0(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t|T_i \geq t, X_i = 00)}{\Delta t} \\ \lambda_1(t) &= \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t|T_i \geq t, X_i \neq 00)}{\Delta t}\end{aligned}$$

Then,

$$\begin{aligned}\mathbb{P}(t \leq T_i < t + \Delta t | T_i \geq t, \mathrm{FH}) &= \sum_{X_i} \mathbb{P}(t \leq T_i < t + \Delta t, X_i | T_i \geq t, \mathrm{FH}) \\ &= \sum_{X_i} \mathbb{P}(t \leq T_i < t + \Delta t, X_i, \mathrm{FH}) / \mathbb{P}(T_i \geq t, \mathrm{FH}) \\ &= \sum_{X_i} \mathbb{P}(t \leq T_i < t + \Delta t | X_i) \mathbb{P}(X_i | T_i \geq t, \mathrm{FH}),\end{aligned}$$

using Bayes' rule and the fact that $\mathbb{P}(t \leq T_i < t + \Delta t | X_i, \mathrm{FH}) = \mathbb{P}(t \leq T_i < t + \Delta t | X_i)$ and $\mathbb{P}(X_i, \mathrm{FH} | T_i \geq t, \mathrm{FH}) = \mathbb{P}(X_i | T_i \geq t, \mathrm{FH})$. Dividing by Δt and taking the limit as Δt tends to 0 gives

$$\lambda(t) = \lambda_1(t) \times \mathbb{P}(X_i \neq 00 | T_i \geq t, \mathrm{FH}) + \lambda_0(t) \times \mathbb{P}(X_i = 00 | T_i \geq t, \mathrm{FH})$$

We showed previously that $\mathbb{P}(X_i \neq 00 | T_i \geq t, \mathrm{FH}) = \pi S_1(t) / (S(t) S_1(\tau))$ and $\mathbb{P}(X_i = 00 | T_i \geq t, \mathrm{FH}) = (1 - \pi) S_0(t) / (S(t) S_0(\tau))$ which concludes the proof. □

Proof of Lemma 3. The first part of the equality is a standard result in the competing risk setting: we have, from Bayes' rule,

$$\lambda_\alpha(u) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T_i < t + \Delta t | \mathrm{FH})}{\Delta t \, \mathbb{P}(T_i^* \geq t | \mathrm{FH})}$$

and consequently $\lambda_\alpha(u) S_\beta(u)$ is equal to the density of T conditionally to FH. Then, since $\lambda_\alpha(u) = \alpha_j$ for $u \in]c_{j-1}, c_j]$ we have

$$\begin{aligned}\mathbb{P}(T_i \leq t | T_i > c_{j-1}, \mathrm{FH}) &= \int_{c_{j-1}}^{t} \lambda_\alpha(u) S_\beta(u) du = \alpha_j \int_{c_{j-1}}^{t} S_\beta(u) du \\ &= \alpha_j \int_{c_{j-1}}^{t} \exp\left(- \int_0^u \lambda_\beta(v) dv \right) du\end{aligned}$$

Now, for $u \in]c_{j-1}, t]$, $t \leq c_j$,

$$\int_0^u \lambda_\beta(v) dv = \int_0^{c_{j-1}} \lambda_\beta(v) dv + \beta_j (u - c_{j-1})$$

and

$$\begin{aligned}\int_{c_{j-1}}^{t} \exp\left(- \int_0^u \lambda_\beta(v) dv \right) du &= \exp\left(- \int_0^{c_{j-1}} \lambda_\beta(v) dv \right) \int_{c_{j-1}}^{t} \exp(-\beta_j (u - c_{j-1})) du \\ &= S_\beta(c_{j-1}) \int_{c_{j-1}}^{t} \exp(-\beta_j (u - c_{j-1})) du\end{aligned}$$

The integral on the right side of the equation is straightforward to compute. This gives,

$$S_\beta(c_{j-1}) \int_{c_{j-1}}^{t} \exp(-\beta_j(u - c_{j-1}))du = \frac{1}{\beta_j}\Big(S_\beta(c_{j-1}) - S_\beta(c_{j-1}) \exp(-\beta_j(t - c_{j-1}))\Big)$$

Finally, we conclude by noticing that

$$\begin{aligned} S_\beta(t) &= \exp\left(-\int_0^{c_{j-1}} \lambda_\beta(u)du - \int_{c_{j-1}}^{t} \lambda_\beta(u)du\right) \\ &= S_\beta(c_{j-1}) \exp\left(-\beta_j(t - c_{j-1})\right) \end{aligned}$$

□

References

Flora Alarcon, Catherine Bourgain, Marion Gauthier-Villars, Violaine Planté-Bordeneuve, D Stoppa-Lyonnet, and Catherine Bonaïti-Pellié. Pel: an unbiased method for estimating age-dependent genetic disease risk from pedigree data unselected for family history. *Genetic epidemiology*, 33(5):379–385, 2009.

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. *Statistical models based on counting processes*. Springer Series in Statistics. Springer-Verlag, New York, 1993. ISBN 0-387-97872-0.

Per Kragh Andersen and Niels Keiding. Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine*, 31 (11-12):1074–1088, 2012.

AC Antoniou, PDP Pharoah, G. McMullan, NE Day, MR Stratton, J. Peto, BJ Ponder, and DF Easton. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. *British journal of cancer*, 86(1):76–83, 2002. ISSN 0007-0920.

AC Antoniou, PPD Pharoah, P. Smith, and DF Easton. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *British journal of cancer*, 91(8):1580–1590, 2004. ISSN 0007-0920.

Sarah D Berry, Long Ngo, Elizabeth J Samelson, and Douglas P Kiel. Competing risk of death: an important consideration in studies of older adults. *Journal of the American Geriatrics Society*, 58(4):783–787, 2010.

Sining Chen, Wenyi Wang, Shing Lee, Khedoudja Nafa, Johanna Lee, Kathy Romans, Patrice Watson, Stephen B Gruber, David Euhus, Kenneth W Kinzler, et al. Prediction of germline mutations and cancer risk in the lynch syndrome. *Jama*, 296(12):1479–1487, 2006.

Elisabeth B Claus, N Risch, and W Douglas Thompson. Genetic analysis of breast cancer in the cancer and steroid hormone study. *American journal of human genetics*, 48(2):232, 1991.

Elizabeth B Claus, Neil Risch, and W Douglas Thompson. Autosomal dominant inheritance of early-onset breast cancer. implications for risk prediction. *Cancer*, 73(3):643–651, 1994.

Adnan Darwiche. Recursive conditioning. *Artificial Intelligence*, 126(1-2):5–41, 2001.

Antoine De Pauw. *Estimation des risques de cancer du sein et de l'ovaire des femmes sans mutation des gènes BRCA1et BRCA2: apport des modèles de calcul de risque*. PhD thesis, Paris 7, 2012.

DF Easton, DT Bishop, D Ford, and GP Crockford. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. the breast cancer linkage consortium. *American journal of human genetics*, 52(4):678, 1993.

Robert C Elston, Varghese T George, and Forrestt Severtson. The elston-stewart algorithm for continuous genotypes and environmental factors. *Human heredity*, 42(1):16–27, 1992.

Maáyan Fishelson and Dan Geiger. Exact genetic linkage computations for general pedigrees. *Bioinformatics*, 18(suppl 1):S189–S198, 2002.

J.M. Hall, M.K. Lee, B. Newman, J.E. Morrow, L.A. Anderson, B. Huey, and M.C. King. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988):1684, 1990. ISSN 0036-8075.

INED. Death incidence in France, period 2012-2014. `https://www.ined.fr/en/everything_about_population/data/france/deaths-causes-mortality/mortality-tables/`, 2017.

Hélène Jacqmin-Gadda, Paul Blanche, Emilie Chary, Lucie Loubère, Hélène Amieva, and Jean-François Dartigues. Prognostic score for predicting risk of dementia over 10 years while accounting for competing risk of death. *American journal of epidemiology*, 180(8):790–798, 2014.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Leonid Kruglyak, Mark J Daly, Mary Pat Reeve-Daly, and Eric S Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. *American journal of human genetics*, 58(6):1347, 1996.

Kenneth Lange, Jeanette C Papp, Janet S Sinsheimer, Ram Sripracha, Hua Zhou, and Eric M Sobel. Mendel: the swiss army knife of genetic analysis programs. *Bioinformatics*, 29(12):1568–1570, 2013.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Steffen L Lauritzen and Nuala A Sheehan. Graphical models for genetic analyses. *Statistical Science*, pages 489–514, 2003.

Amir Mehrgou and Mansoureh Akouchekian. The importance of brca1 and brca2 genes mutations in breast cancer development. *Medical Journal of the Islamic Republic of Iran*, 30:369, 2016.

Jeffrey R O'Connell and Daniel E Weeks. Pedcheck: a program for identification of genotype incompatibilities in linkage analysis. *The American Journal of Human Genetics*, 63(1):259–266, 1998.

Kimmo Palin, Harry Campbell, Alan F Wright, James F Wilson, and Richard Durbin. Identity-by-descent-based phasing and imputation in founder populations using graphical models. *Genetic epidemiology*, 35(8):853–860, 2011.

V. Plante-Bordeneuve, J. Carayol, A. Ferreira, D. Adams, F. Clerget-Darpoux, M. Misrahi, G. Said, and C. Bonaïti-Pellié. Genetic study of transthyretin amyloid neuropathies: carrier risks among French and Portuguese families. *J Med Genet*, 40(11):e120, 2003.

Liviu R Totir, Rohan L Fernando, and Joseph Abraham. An efficient algorithm to compute marginal posterior genotype probabilities for every member of a pedigree with loops. *Genetics Selection Evolution*, 41(1):52, 2009.

Mathilde Wanneveich, Hélène Jacqmin-Gadda, Jean-François Dartigues, and Pierre Joly. Impact of intervention targeting risk factors on chronic disease burden. *Statistical methods in medical research*, page 0962280216631360, 2016.

Bibliography

Odd Aalen. Nonparametric inference for a family of counting processes. The Annals of Statistics, pages 701–726, 1978.

Odd Aalen, Ornulf Borgan, and Hakon Gjessing. Survival and event history analysis: a process point of view. Springer Science & Business Media, 2008.

Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. Knowledge and information systems, 51(2):339–367, 2017.

Per Kragh Andersen. Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. Statistics in medicine, 7(6):661–670, 1988.

Thierry André, Armand De Gramont, Dewi Vernerey, Benoist Chibaudel, Franck Bonnetain, Annemilaï Tijeras-Raballand, Aurelie Scriva, Tamas Hickish, Josep Tabernero, Jean Luc Van Laethem, et al. Adjuvant fluorouracil, leucovorin, and oxaliplatin in stage II to III colon cancer: updated 10-year survival and outcomes according to BRAF mutation and mismatch repair status of the MOSAIC study. Journal of Clinical Oncology, 33(35):4176–4187, 2015.

Antonis C Antoniou, Paul DP Pharoah, Greg McMullan, Nickolas E Day, MR Stratton, J Peto, BJ Ponder, and Douglas F Easton. A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. British journal of cancer, 86(1):76–83, 2002.

Antonis C Antoniou, Paul DP Pharoah, Paula Smith, and Douglas F Easton. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. British journal of cancer, 91(8):1580–1590, 2004.

Antonis C Antoniou, AP Cunningham, J Peto, DG Evans, F Lalloo, SA Narod, HA Risch, JE Eyfjord, JL Hopper, MC Southey, et al. The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. British journal of cancer, 98(8):1457–1466, 2008.

Stefan Arnborg, Derek G Corneil, and Andrzej Proskurowski. Complexity of finding embeddings in a k-tree. SIAM Journal on Algebraic Discrete Methods, 8(2):277–284, 1987.

Richard Arratia, Larry Goldstein, and Louis Gordon. Poisson approximation and the Chen-Stein method. Statistical Science, pages 403–424, 1990.

Nazila Assasi, G Blackhouse, K Campbell, K Gaebel, R Hopkins, J Jegathisawaran, A Sinclair, K Seal, C Kamel, M Levine, et al. DNA Mismatch repair deficiency tumour testing for patients with colorectal cancer: a health technology assessment. 2016.

John AD Aston and Donald EK Martin. Distributions associated with general runs and patterns in hidden Markov models. The Annals of Applied Statistics, 1(2): 585–611, 2007.

Ivan E Auger and Charles E Lawrence. Algorithms for the optimal identification of segment neighborhoods. Bulletin of mathematical biology, 51(1):39–54, 1989.

Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. The Journal of experimental medicine, 79(2):137–158, 1944.

Rebecca A Barnetson, Albert Tenesa, Susan M Farrington, Iain D Nicholl, Roseanne Cetnarskyj, Mary E Porteous, Harry Campbell, and Malcolm G Dunlop. Identification and survival of carriers of mutations in DNA mismatch-repair genes in colon cancer. New England Journal of Medicine, 354(26):2751–2763, 2006.

Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The annals of mathematical statistics, 41(1):164–171, 1970.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. Journal of machine learning research, 18, 2018.

Ann Becker, Dan Geiger, and Alejandro A Schäffer. Automatic Selection of Loop Breakers for Genetic Linkage Analysis. Human heredity, 48(1):49–60, 1998.

Patrick R Benusiglio, Florence Coulet, Alexandra Lefebvre, Alex Duval, and Grégory Nuel. Overcoming the challenges associated with universal screening for Lynch syndrome in colorectal and endometrial cancer. Genetics in Medicine, 22(8):1422–1423, 2020.

Donald A Berry, Giovanni Parmigiani, Juana Sanchez, Joellen Schildkraut, and Eric Winer. Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history. Journal of the National Cancer Institute, 89(3):227–237, 1997.

Patrick M Boland, Matthew B Yurgelun, and C Richard Boland. Recent progress in Lynch syndrome and other familial colorectal cancer syndromes. CA: a cancer journal for clinicians, 68(3):217–231, 2018.

Maxime Bonhomme, Maria Inés Fariello, Hélène Navier, Ahmed Hajri, Yacine Badis, Henri Miteul, Deborah A Samac, Bernard Dumas, Alain Baranger, Christophe Jacquet, et al. A local score approach improves GWAS resolution and detects minor QTL: application to Medicago truncatula quantitative disease resistance to multiple Aphanomyces euteiches isolates. Heredity, 123(4):517–531, 2019.

Theodor Boveri. Ergebnisse uber die Konstitution der chromatischen Substanz des Zellkerns. . Jena,G. Fischer,, 1904. URL https://www.biodiversitylibrary.org/item/70637. https://www.biodiversitylibrary.org/bibliography/28064.

Jerome V Braun and Hans-Georg Muller. Statistical methods for DNA sequence segmentation. Statistical Science, pages 142–162, 1998.

Adam R Brentnall, Wendy F Cohn, William A Knaus, Martin J Yaffe, Jack Cuzick, and Jennifer A Harvey. A case-control study to add volumetric or clinical mammographic density into the Tyrer-Cuzick breast cancer risk model. Journal of breast imaging, 1(2):99–106, 2019.

Norman E Breslow. Contribution to discussion of paper by DR Cox. J. Roy. Statist. Soc., Ser. B, 34:216–217, 1972.

Norman E Breslow. Covariance analysis of censored survival data. Biometrics, pages 89–99, 1974.

Bruno Buecher, Marine Le Mentec, François Doz, Franck Bourdeaut, Marion Gauthier-Villars, Dominique Stoppa-Lyonnet, and Chrystelle Colas. Syndrome CMMRD (déficience constitutionnelle des gènes MMR): bases génétiques et aspects cliniques. Bulletin du Cancer, 106(2):162–172, 2019.

John Burn, Anne-Marie Gerdes, Finlay Macrae, Jukka-Pekka Mecklin, Gabriela Moeslein, Sylviane Olschwang, Diane Eccles, D Gareth Evans, Eamonn R Maher, Lucio Bertario, et al. Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial. The Lancet, 378(9809):2081–2087, 2011.

Guillaume Canard, Jeremie H Lefevre, Chrystelle Colas, Florence Coulet, Magali Svrcek, Olivier Lascols, Richard Hamelin, Conor Shields, Alex Duval, Jean-Francois Fléjou, et al. Screening for Lynch syndrome in colorectal cancer: are we doing enough? Annals of surgical oncology, 19(3):809–816, 2012.

Chris Cannings, Elizabeth A Thompson, and Mark H Skolnick. The recursive derivation of likelihoods on complex pedigrees. Advances in Applied Probability, 8(4): 622–625, 1976.

Olivier Cappé and Eric Moulines. Recursive computation of the score and observed information matrix in hidden Markov models. In IEEE/SP 13th Workshop on Statistical Signal Processing, 2005, pages 703–708. IEEE, 2005.

Olivier Cappé, Eric Moulines, and Tobias Rydén. Inference in hidden Markov models. Springer Science & Business Media, 2005.

Rita Casadio, Piero Fariselli, Chiara Taroni, and Mario Compiani. A predictor of transmembrane α-helix domains of proteins based on neural networks. European biophysics journal, 24(3):165–178, 1996.

Enrique Castillo, José Manuel Gutiérrez, and Ali S Hadi. Sensitivity analysis in discrete Bayesian networks. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 27(4):412–423, 1997.

Guia Cerretelli, Ann Ager, Mark J Arends, and Ian M Frayling. Molecular pathology of Lynch syndrome. The Journal of pathology, 250(5):518–531, 2020.

Hei Chan and Adnan Darwiche. When do numbers really matter? Journal of artificial intelligence research, 17:265–287, 2002.

Hei Chan and Adnan Darwiche. Sensitivity analysis in Bayesian networks: From single to multiple parameters. arXiv preprint arXiv:1207.4124, 2012.

Hsiu-Hsi Chen, Stephen W Duffy, and Laszlo Tabar. A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. Journal of the Royal Statistical Society: Series D (The Statistician), 45(3):307–317, 1996.

Jie Chen and Joseph Glaz. Scan statistics for monitoring data modeled by a negative binomial distribution. Communications in Statistics-Theory and Methods, 45(6): 1632–1642, 2016.

Sining Chen and Giovanni Parmigiani. Meta-analysis of BRCA1 and BRCA2 penetrance. Journal of clinical oncology: official journal of the American Society of Clinical Oncology, 25(11):1329, 2007.

Sining Chen, Wenyi Wang, Shing Lee, Khedoudja Nafa, Johanna Lee, Kathy Romans, Patrice Watson, Stephen B Gruber, David Euhus, Kenneth W Kinzler, et al. Prediction of germline mutations and cancer risk in the Lynch syndrome. Jama, 296(12):1479–1487, 2006.

Michael Herman Chui, Paul Ryan, Jordan Radigan, Sarah E Ferguson, Aaron Pollett, Melyssa Aronson, Kara Semotiuk, Spring Holter, Keiyan Sy, Janice S Kwon, et al. The histomorphology of Lynch syndrome–associated ovarian carcinomas: toward a subtype-specific screening strategy. The American journal of surgical pathology, 38(9):1173–1181, 2014.

Jessica A Cintolo-Gonzalez, Danielle Braun, Amanda L Blackford, Emanuele Mazzola, Ahmet Acar, Jennifer K Plichta, Molly Griffin, and Kevin S Hughes. Breast cancer risk models: a comprehensive overview of existing models, validation, and clinical applications. Breast cancer research and treatment, 164(2):263–284, 2017.

Elisabeth B Claus, Neil Risch, and W Douglas Thompson. Genetic analysis of breast cancer in the cancer and steroid hormone study. American journal of human genetics, 48(2):232, 1991.

Elizabeth B Claus, Neil Risch, and W Douglas Thompson. Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction. Cancer, 73(3):643–651, 1994.

Raphaël Colle, Romain Cohen, Delphine Cochereau, Alex Duval, Olivier Lascols, Daniel Lopez-Trabada, Pauline Afchain, Isabelle Trouilloud, Yann Parc, Jérémie H Lefevre, et al. Immunotherapy and patients treated for cancer with microsatellite instability. Bulletin du cancer, 104(1):42–51, 2017.

Daniel Commenges. Multi-state models in epidemiology. Lifetime data analysis, 5 (4):315–327, 1999.

Gregory F Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. Artificial intelligence, 42(2-3):393–405, 1990.

Robert W Cottingham Jr, Ramana M Idury, and Alejandro A Schäffer. Faster sequential genetic linkage computations. American journal of human genetics, 53 (1):252, 1993.

Robert G Cowell. Calculating moments of decomposable functions in Bayesian networks. Research Report 109, Department of Statistical Sciences, University College London, London., 1992.

Robert G Cowell, Philip Dawid, Steffen L Lauritzen, and David J Spiegelhalter. Probabilistic networks and expert systems: Exact computational methods for Bayesian networks. Springer Science & Business Media, 1999.

Emma J Crosbie, Neil AJ Ryan, Mark J Arends, Tjalling Bosse, John Burn, Joanna M Cornes, Robin Crawford, Diana Eccles, Ian M Frayling, Sadaf Ghaem-Maghami, et al. The Manchester International Consensus Group recommendations for the management of gynecological cancers in Lynch syndrome. Genetics in Medicine, 21(10):2390–2400, 2019.

Emma J Crosbie, Neil AJ Ryan, Rhona J McVey, Fiona Lalloo, Naomi Bowers, Kate Green, Emma R Woodward, Tara Clancy, James Bolton, Andrew J Wallace, et al. Assessment of mismatch repair deficiency in ovarian cancer. Journal of Medical Genetics, 2020.

Adnan Darwiche. A differential approach to inference in Bayesian networks. Journal of the ACM (JACM), 50(3):280–305, 2003.

Ghazaleh S Dashti, Wing Yan Li, Daniel D Buchanan, Mark Clendenning, Christophe Rosty, Ingrid M Winship, Finlay A Macrae, Graham G Giles, Sheetal Hardikar, Xinwei Hua, et al. Type 2 diabetes mellitus, blood cholesterol, triglyceride and colorectal cancer risk in Lynch syndrome. British journal of cancer, pages 1–8, 2019.

Albert de la Chapelle, Glenn Palomaki, and Heather Hampel. Identifying lynch syndrome. International journal of cancer. Journal international du cancer, 125 (6):1492, 2009.

Wiljo JF de Leeuw, JanWillem Dierssen, Hans FA Vasen, Juul Th Wijnen, Gemma G Kenter, Hanne Meijers-Heijboer, Annette Brocker-Vriends, Astrid Stormorken, Pal Moller, Fred Menko, et al. Prediction of a mismatch repair gene defect by microsatellite instability and immunohistochemical analysis in endometrial tumours from HNPCC patients. The Journal of pathology, 192(3):328–335, 2000.

Antoine de Pauw. Estimation des risques de cancer du sein et de l'ovaire des femmes sans mutation des gène PhD thesis, Paris 7, 2012.

Thomas A Dean, Sumeetpal S Singh, Ajay Jasra, and Gareth W Peters. Parameter estimation for hidden Markov models with intractable likelihoods. Scandinavian Journal of Statistics, 41(4):970–987, 2014.

Gautier Defossez, Sandra Le Guyader-Peyrou, Zoé Uhry, Pascale Grosclaude, Marc Colonna, Emmanuelle Dantony, Patricia Delafosse, Florence Molinié, Anne-Sophie Woronoff, Anne-Marie Bouvier, et al. Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018. Etude à partir des registres des cancers du réseau Francim. Résultats préliminaires. Synthèse. Saint-Maurice (Fra): Santé Publique France, page 19, 2019.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

Ivan Diaz-Padilla, Nuria Romero, Eitan Amir, Xavier Matias-Guiu, Eduardo Vilar, Franco Muggia, and Jesus Garcia-Donas. Mismatch repair status and clinical outcome in endometrial cancer: a systematic review and meta-analysis. Critical reviews in oncology/hematology, 88(1):154–167, 2013.

Mev Dominguez-Valentin, Emma J Crosbie, Christoph Engel, Stefan Aretz, Finlay Macrae, Ingrid Winship, Gabriel Capella, Huw Thomas, Sigve Nakken, Eivind Hovig, et al. Risk-reducing hysterectomy and bilateral salpingo-oophorectomy in female heterozygotes of pathogenic mismatch repair variants: a Prospective Lynch Syndrome Database report. Genetics in Medicine, pages 1–8, 2020a.

Mev Dominguez-Valentin, Julian R Sampson, Toni T Seppälä, Sanne W Ten Broeke, John-Paul Plazzer, Sigve Nakken, Christoph Engel, Stefan Aretz, Mark A Jenkins, Lone Sunde, et al. Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database. Genetics in Medicine, 22(1), 2020b.

Coralie Dorard, Aurélie De Thonel, Ada Collura, Laetitia Marisa, Magali Svrcek, Anaïs Lagrange, Gaetan Jego, Kristell Wanherdrick, Anne Laure Joly, Olivier Buhard, et al. Expression of a mutant HSP110 sensitizes colorectal cancer cells to

chemotherapy and improves disease prognosis. Nature medicine, 17(10):1283–1289, 2011.

Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press, 1998.

Douglas F Easton, DT Bishop, D Ford, and GP Crockford. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. American journal of human genetics, 52(4):678, 1993.

Douglas F Easton, Amie M Deffenbaugh, Dmitry Pruss, Cynthia Frye, Richard J Wenstrup, Kristina Allen-Brady, Sean V Tavtigian, Alvaro NA Monteiro, Edwin S Iversen, Fergus J Couch, et al. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer–predisposition genes. The American Journal of Human Genetics, 81(5):873–883, 2007.

Robert C Elston. Ascertainment and age of onset in pedigree analysis. Human heredity, 23(2):105–112, 1973.

Robert C Elston and John Stewart. A general model for the genetic analysis of pedigree data. Human heredity, 21(6):523–542, 1971.

David Enard, Philipp W Messer, and Dmitri A Petrov. Genome-wide signals of positive selection in human evolution. Genome research, 24(6):885–895, 2014.

Manel Esteller, Ross Levine, Stephen B Baylin, Lora Hedrick Ellenson, and James G Herman. MLH1 promoter hypermethylation is associated with the microsatellite instability phenotype in sporadic endometrial carcinomas. Oncogene, 17(18):2413–2417, 1998.

María Inés Fariello, Simon Boitard, Sabine Mercier, David Robelin, Thomas Faraut, Cécile Arnould, Julien Recoquillay, Olivier Bouchez, Gérald Salin, Patrice Dehais, et al. Accounting for linkage disequilibrium in genome scans for selection without individual genotypes: the local score approach. Molecular ecology, 26(14):3700–3714, 2017.

Maayan Fishelson and Dan Geiger. Exact genetic linkage computations for general pedigrees. Bioinformatics, 18(suppl_1):S189–S198, 2002.

Ronald A Fisher. XV. The correlation between relatives on the supposition of Mendelian inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 52(2):399–433, 1919.

Ronald Aylmer Fisher. The genetical theory of natural selection. Clarendon Press, 1930.

George David Forney. The Viterbi algorithm. Proceedings of the IEEE, 61(3):268–278, 1973.

Yohann Foucher, Eve Mathieu, Philippe Saint-Pierre, Jean-François Durand, and Jean-Pierre Daurès. A semi-Markov model based on generalized Weibull distribution with an illustration for HIV disease. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 47(6):825–833, 2005.

Jane Fridlyand, Antoine M Snijders, Dan Pinkel, Donna G Albertson, and Ajay N Jain. Hidden Markov models approach to the analysis of array CGH data. Journal of multivariate analysis, 90(1):132–153, 2004.

James C Fu and Markos V Koutras. Distribution theory of runs: a Markov chain approach. Journal of the American Statistical Association, 89(427):1050–1058, 1994.

James C Fu, Liqun Wang, and WY Wendy Lou. On exact and large deviation approximation for the distribution of the longest run in a sequence of two-state Markov dependent trials. Journal of Applied Probability, 40(2):346–360, 2003.

Mitchell H Gail. Twenty-five years of breast cancer risk models and their applications. JNCI: Journal of the National Cancer Institute, 107(5), 2015.

Mitchell H Gail, Louise A Brinton, David P Byar, Donald K Corle, Sylvan B Green, Catherine Schairer, and John J Mulvihill. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. JNCI: Journal of the National Cancer Institute, 81(24):1879–1886, 1989.

Mitchell H Gail, Joseph P Costantino, David Pee, Melissa Bondy, Lisa Newman, Mano Selvan, Garnet L Anderson, Kathleen E Malone, Polly A Marchbanks, Worta McCaskill-Stevens, et al. Projecting individualized absolute invasive breast cancer risk in African American women. Journal of the National Cancer Institute, 99(23): 1782–1792, 2007.

RC Gentleman, JF Lawless, JC Lindsey, and P Yan. Multi-state Markov models for analysing incomplete disease history data with illustrations for HIV disease. Statistics in medicine, 13(8):805–821, 1994.

Mark X Geske, Anant P Godbole, Andrew A Schaffner, Allison M Skolnick, and Garrick L Wallstrom. Compound Poisson approximations for word patterns under Markovian hypotheses. Journal of applied probability, 32(4):877–892, 1995.

Joseph Glaz, Vladimir Pozdnyakov, and Sylvan Wallenstein. Scan statistics: Methods and applications. Springer Science & Business Media, 2009.

David E Goldgar, Douglas F Easton, Amie M Deffenbaugh, Alvaro NA Monteiro, Sean V Tavtigian, Fergus J Couch, Breast Cancer Information Core (BIC) Steering Committee, et al. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. The American Journal of Human Genetics, 75(4):535–544, 2004.

David E Goldgar, Douglas F Easton, Graham B Byrnes, Amanda B Spurdle, Edwin S Iversen, Marc S Greenblatt, and IARC Unclassified Genetic Variants Working Group. Genetic evidence and integration of various data sources for classifying uncertain variants into a single model. Human mutation, 29(11):1265–1272, 2008.

Paul J Goodfellow, Caroline C Billingsley, Heather A Lankes, Shamshad Ali, David E Cohn, Russell J Broaddus, Nilsa Ramirez, Colin C Pritchard, Heather Hampel, Alexis S Chassen, et al. Combined microsatellite instability, MLH1 methylation analysis, and immunohistochemistry for Lynch syndrome screening in endometrial cancers from GOG210: an NRG Oncology and Gynecologic Oncology Group study. Journal of Clinical Oncology, 33(36):4301, 2015.

Malka Gorfine, Li Hsu, and Giovanni Parmigiani. Frailty models for familial risk with application to breast cancer. Journal of the American Statistical Association, 108 (504):1205–1215, 2013.

Sharon R Grossman, Kristian G Andersen, Ilya Shlyakhter, Shervin Tabrizi, Sarah Winnicki, Angela Yen, Daniel J Park, Dustin Griesemer, Elinor K Karlsson, Sunny H Wong, et al. Identifying recent adaptations in large-scale genomic data. Cell, 152(4):703–713, 2013.

Daniel F Gudbjartsson, Kristjan Jonasson, Michael L Frigge, and Augustine Kong. Allegro, a new computer program for multipoint linkage analysis. Nature genetics, 25(1):12–13, 2000.

Yann Guédon. Exploring the state sequence space for hidden Markov and semi-Markov chains. Computational Statistics & Data Analysis, 51(5):2379–2409, 2007.

Laurent Guéguen. Sarment: Python modules for HMM analysis and partitioning of sequences. Bioinformatics, 21(16):3427–3428, 2005.

Chantal Guihenneuc-Jouyaux, Sylvia Richardson, and Ira M Longini Jr. Modeling markers of disease progression by a hidden Markov process: application to characterizing CD4 cell decline. Biometrics, 56(3):733–741, 2000.

Hanmin Guo, James J Li, Qiongshi Lu, and Lin Hou. Detecting local genetic correlations with scan statistics. Nature communications, 12(1):1–13, 2021.

John BS Haldane. The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet, 8(29):299–309, 1919.

Heather Hampel, Wendy L Frankel, Edward Martin, Mark Arnold, Karamjit Khanduja, Philip Kuebler, Mark Clendenning, Kaisa Sotamaa, Thomas Prior, Judith A Westman, et al. Feasibility of screening for Lynch syndrome among patients with colorectal cancer. Journal of Clinical Oncology, 26(35):5783, 2008.

Heather Hampel, Rachel Pearlman, Mallory Beightol, Weiqiang Zhao, Daniel Jones, Wendy L Frankel, Paul J Goodfellow, Ahmet Yilmaz, Kristin Miller, Jason Bacher, et al. Assessment of tumor sequencing as a replacement for Lynch syndrome

screening and current molecular tests for patients with colorectal cancer. JAMA oncology, 4(6):806–813, 2018.

Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. cell, 100(1): 57–70, 2000.

Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. cell, 144(5):646–674, 2011.

Zihuai He, Bin Xu, Joseph Buxbaum, and Iuliana Ionita-Laza. A genome-wide scan statistic framework for whole-genome sequence data analysis. Nature communications, 10(1):1–11, 2019.

Yvonne MC Hendriks, Anja Wagner, Hans Morreau, Fred Menko, Astrid Stormorken, Franz Quehenberger, Lodewijk Sandkuijl, Pal Møller, Maurizio Genuardi, Hans Van Houwelingen, et al. Cancer risk in hereditary nonpolyposis colorectal cancer due to MSH6 mutations: impact on counseling and surveillance. Gastroenterology, 127(1):17–25, 2004.

Megan P Hitchins, Justin JL Wong, Graeme Suthers, Catherine M Suter, David IK Martin, Nicholas J Hawkins, and Robyn L Ward. Inheritance of a cancer-associated MLH1 germ-line epimutation. New England Journal of Medicine, 356(7):697–705, 2007.

Toby Dylan Hocking, Guillem Rigaill, Paul Fearnhead, and Guillaume Bourque. Generalized functional pruning optimal partitioning (gfpop) for constrained change-point detection in genomic data. arXiv preprint arXiv:1810.00117, 2018.

Philip Hougaard. Frailty models for survival data. Lifetime data analysis, 1(3): 255–273, 1995.

Philip Hougaard. Multi-state models: a review. Lifetime data analysis, 5(3):239–264, 1999.

Philip Hougaard. Analysis of multivariate survival data. Springer Science & Business Media, 2012.

Christopher Jackson and Maintainer Christopher Jackson. Package "msm", 2019.

Christopher H Jackson and Linda D Sharples. Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. Statistics in medicine, 21(1):113–128, 2002.

Christopher H Jackson, Linda D Sharples, Simon G Thompson, Stephen W Duffy, and Elisabeth Couto. Multistate Markov models for disease progression with classification error. Journal of the Royal Statistical Society: Series D (The Statistician), 52(2):193–209, 2003.

Christopher H Jackson et al. Multi-state models for panel data: the msm package for R. Journal of statistical software, 38(8):1–29, 2011.

Heikki J Järvinen, Markku Aarnio, Harri Mustonen, Katja Aktan-Collan, Lauri A Aaltonen, Päivi Peltomäki, Albert De La Chapelle, and Jukka-Pekka Mecklin. Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer. Gastroenterology, 118(5):829–834, 2000.

Claus Skaanning Jensen and Augustine Kong. Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. The American Journal of Human Genetics, 65(3):885–901, 1999.

Finn V Jensen and Thomas Dyhre Nielsen. Bayesian networks and decision graphs, volume 2. Springer, 2007.

Michael F Kane, Massimo Loda, Gretchen M Gaida, Jennifer Lipman, Rajesh Mishra, Harvey Goldman, J Milburn Jessup, and Richard Kolodner. Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. Cancer research, 57(5):808–811, 1997.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282):457–481, 1958.

Samuel Karlin. Statistical signals in bioinformatics. Proceedings of the National Academy of Sciences, 102(38):13355–13362, 2005.

Samuel. Karlin and Stephen F Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences, 87(6):2264–2268, 1990.

Fay Kastrinos, Hajime Uno, Chinedu Ukaegbu, Carmelita Alvero, Ashley McFarland, Matthew B Yurgelun, Matthew H Kulke, Deborah Schrag, Jeffrey A Meyerhardt, Charles S Fuchs, et al. Development and validation of the PREMM5 model for comprehensive risk assessment of Lynch syndrome. Journal of Clinical Oncology, 35(19):2165, 2017.

Richard Kay. A Markov model for analysing cancer markers and disease states in survival studies. Biometrics, pages 855–865, 1986.

Zohreh Ketabi, Katarina Bartuma, Inge Bernstein, Susanne Malander, Henrik Grönberg, Erik Björck, Susanne Holck, and Mef Nilbert. Ovarian cancer linked to Lynch syndrome typically presents as early-onset, non-serous epithelial tumors. Gynecologic oncology, 121(3):462–465, 2011.

Kenneth W Kinzler and Bert Vogelstein. Gatekeepers and caretakers. Nature, 386 (6627):761–763, 1997.

Uffe Kjærulff. Triangulation of graphs–algorithms giving small total state space. 1990.

Daphne Koller and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

Augustine Kong, DC Rao, and GP Vogler. Efficient methods for computing linkage likelihoods of recessive diseases in inbred pedigrees. Genetic Epidemiology, 8(2): 81–103, 1991.

Anders Krogh, Björn Larsson, Gunnar Von Heijne, and Erik LL Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology, 305(3):567–580, 2001.

Leonid Kruglyak, Mark J Daly, Mary Pat Reeve-Daly, and Eric S Lander. Parametric and nonparametric linkage analysis: a unified multipoint approach. American journal of human genetics, 58(6):1347, 1996.

Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. Journal of molecular biology, 157(1):105–132, 1982.

Eric S Lander and Philip Green. Construction of multilocus genetic linkage maps in humans. Proceedings of the National Academy of Sciences, 84(8):2363–2367, 1987.

Eric S Lander, Philip Green, Jeff Abrahamson, Aaron Barlow, Mark J Daly, Stephen E Lincoln, and Lee Newburg. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics, 1(2):174–181, 1987.

K Lange and RC Elston. Extensions to pedigree analysis. Human heredity, 25(2): 95–105, 1975.

Kenneth Lange, Jeanette C. Papp, Janet S. Sinsheimer, R. Sripracha, H. Zhou, and E. M. Sobel. Mendel: the Swiss army knife of genetic analysis programs. Bioinformatics, 29(12):1568–1570, 2013.

Steffen L Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. Journal of the American Statistical Association, 87 (420):1098–1108, 1992.

Steffen L Lauritzen. Graphical models, volume 17. Clarendon Press, 1996.

Steffen L Lauritzen and Nuala A Sheehan. Graphical models for genetic analyses. Statistical Science, pages 489–514, 2003.

Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. Journal of the Royal Statistical Society: Series B (Methodological), 50(2):157–194, 1988.

Noémie Lavoine, Chrystelle Colas, Martine Muleris, Sahra Bodo, Alex Duval, Natacha Entz-Werle, Florence Coulet, O Cabaret, Felipe Andreiuolo, C Charpy, et al. Constitutional mismatch repair deficiency syndrome: clinical description in a French cohort. Journal of medical genetics, 52(11):770–778, 2015.

Dung T Le, Jennifer N Uram, Hao Wang, Bjarne R Bartlett, Holly Kemberling, Aleksandra D Eyring, Andrew D Skora, Brandon S Luber, Nilofer S Azad, Dan Laheru, et al. PD-1 blockade in tumors with mismatch-repair deficiency. New England Journal of Medicine, 372(26):2509–2520, 2015.

Andrew Lee, Nasim Mavaddat, Amber N Wilcox, Alex P Cunningham, Tim Carver, Simon Hartley, Chantal Babb de Villiers, Angel Izquierdo, Jacques Simard, Marjanka K Schmidt, et al. BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. Genetics in Medicine, 21 (8):1708–1718, 2019.

Andrew J Lee, Alex P Cunningham, KB Kuchenbaecker, N Mavaddat, Douglas F Easton, and Antonis C Antoniou. BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. British journal of cancer, 110(2):535–545, 2014.

Alexandra Lefebvre and Grégory Nuel. A sum-product algorithm with polynomials for computing exact derivatives of the likelihood in Bayesian networks. In Proceedings of Machine Learning Research, International Conference on Probabilistic Graphical Models, pages 201–212, 2018.

Marjolijn JL Ligtenberg, Roland P Kuiper, Tsun Leung Chan, Monique Goossens, Konnie M Hebeda, Marsha Voorendt, Tracy YH Lee, Danielle Bodmer, Eveline Hoenselaar, Sandra JB Hendriks-Cornelissen, et al. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. Nature genetics, 41(1):112–117, 2009.

David V Lindley. The theory of queues with a single server. In Mathematical Proceedings of the Cambridge Philosophical Society, volume 48, pages 277–289. Cambridge University Press, 1952.

Manuel E Lladser. Minimal Markov chain embeddings of pattern problems. 2006. In Proceedings of the 2007 Information Theory and Applications Workshop, University of California, San Diego.

M Lothaire. Applied combinatorics on words, volume 105. Cambridge University Press, 2005.

Thomas A Louis. Finding the observed information matrix when using the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 44 (2):226–233, 1982.

Alexander V Lukashin and Mark Borodovsky. GeneMark. HMM: new solutions for gene finding. Nucleic acids research, 26(4):1107–1115, 1998.

The Minh Luong, Yves Rozenholc, and Grégory Nuel. Fast estimation of posterior probabilities in change-point analysis through a constrained hidden Markov model. Computational Statistics & Data Analysis, 68:129–140, 2013.

Henry T Lynch and Anne J Krush. Cancer family "G" revisited: 1895-1970. Cancer, 27(6):1505–1511, 1971.

Henry T Lynch and Thomas Smyrk. Hereditary nonpolyposis colorectal cancer (Lynch syndrome): an updated review. Cancer: Interdisciplinary International Journal of the American Cancer Society, 78(6):1149–1167, 1996.

Henry T Lynch, W Kimberling, William A Albano, Jane F Lynch, Karen Biscone, Guy S Schuelke, Avery A Sandberg, Lipkin Martin, Eleanor E Deschner, Yves B Mikol, et al. Hereditary nonpolyposis colorectal cancer (Lynch syndromes I and II). I. Clinical description of resource. Cancer, 56(4):934–938, 1985a.

Henry T Lynch, Guy S Schuelke, William J Kimberling, William A Albano, Jane F Lynch, Karen A Biscone, Martin L Lipkin, Eleanor E Deschner, Yves B Mikol, Avery A Sandberg, et al. Hereditary nonpolyposis colorectal cancer (Lynch syndromes I and II). II. Biomarker studies. Cancer, 56(4):939–951, 1985b.

Henry T Lynch, C Richard Boland, Miguel A Rodriguez-Bigas, Christopher Amos, Jane F Lynch, and Patrick M Lynch. Who should be sent for genetic testing in hereditary colorectal cancer syndromes? Journal of Clinical Oncology, 25(23): 3534–3542, 2007.

Paul M Magtibay, Jessica L Nyholm, Jose L Hernandez, and Karl C Podratz. Ovarian remnant syndrome. American journal of obstetrics and gynecology, 193(6):2062–2066, 2005.

Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate Bayesian computational methods. Statistics and Computing, 22(6): 1167–1180, 2012.

John C Marioni, Natalie P Thorne, and Simon Tavaré. BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. Bioinformatics, 22(9): 1144–1146, 2006.

Guillermo Marshall and Richard H Jones. Multi-state models and diabetic retinopathy. Statistics in medicine, 14(18):1975–1983, 1995.

Donald EK Martin and John AD Aston. Distribution of statistics of hidden state sequences through the sum-product algorithm. Methodology and Computing in Applied Probability, 15(4):897–918, 2013.

Rayna K Matsuno, Joseph P Costantino, Regina G Ziegler, Garnet L Anderson, Huilin Li, David Pee, and Mitchell H Gail. Projecting individualized absolute invasive breast cancer risk in Asian and Pacific Islander American women. Journal of the National Cancer Institute, 103(12):951–961, 2011.

Emanuele Mazzola, Jonathan Chipman, Su-Chun Cheng, and Giovanni Parmigiani. Recent BRCAPRO upgrades significantly improve calibration. Cancer Epidemiology and Prevention Biomarkers, 23(8):1689–1695, 2014.

Emanuele Mazzola, Amanda Blackford, Giovanni Parmigiani, and Swati Biswas. Recent enhancements to the genetic risk prediction model BRCAPRO. Cancer informatics, 14:CIN–S17292, 2015.

Anne Marie McCarthy, Zoe Guan, Michaela Welch, Molly E Griffin, Dorothy A Sippo, Zhengyi Deng, Suzanne B Coopey, Ahmet Acar, Alan Semine, Giovanni Parmigiani, et al. Performance of breast cancer risk-assessment models in a large mammography cohort. JNCI: Journal of the National Cancer Institute, 112(5): 489–497, 2020.

MK McConechy, A Talhouk, HH Li-Chang, S Leung, DG Huntsman, CB Gilks, and JN McAlpine. Detection of DNA mismatch repair (MMR) deficiencies by immunohistochemistry can effectively diagnose the microsatellite instability (MSI) phenotype in endometrial carcinomas. Gynecologic oncology, 137(2):306–310, 2015.

Geoff J McLachlan and David C McGiffin. On the role of finite mixture models in survival analysis. Statistical methods in medical research, 3(3):211–226, 1994.

Luís Meira-Machado, Jacobo de Uña-Álvarez, Carmen Cadarso-Suárez, and Per K Andersen. Multi-state models for the analysis of time-to-event data. Statistical methods in medical research, 18(2):195–222, 2009.

Gregor Mendel. Experiments in plant hybridization. Verhandlungen des naturforschenden Vereins Brünn) Available online, 1866.

Arjen R Mensenkamp, Ingrid P Vogelaar, Wendy AG van Zelst-Stams, Monique Goossens, Hicham Ouchene, Sandra JB Hendriks-Cornelissen, Michael P Kwint, Nicoline Hoogerbrugge, Iris D Nagtegaal, and Marjolijn JL Ligtenberg. Somatic mutations in MLH1 and MSH2 are a frequent cause of mismatch-repair deficiency in Lynch syndrome-like tumors. Gastroenterology, 146(3):643–646, 2014.

Sabine Mercier and Grégory Nuel. Duality between the local score of one sequence and constrained Hidden Markov Model. Methodology and Computing in Applied Probability, pages 1–28, 2021.

Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM review, 45(1):3–49, 2003.

Pål Møller, Toni Seppälä, Inge Bernstein, Elke Holinski-Feder, Paola Sala, D Gareth Evans, Annika Lindblom, Finlay Macrae, Ignacio Blanco, Rolf Sijmons, et al. Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database. Gut, 66(3):464–472, 2017a.

Pål Møller, Toni Seppälä, Inge Bernstein, Elke Holinski-Feder, Paola Sala, D Gareth Evans, Annika Lindblom, Finlay Macrae, Ignacio Blanco, Rolf Sijmons, et al. Incidence of and survival after subsequent cancers in carriers of pathogenic MMR variants with previous cancer: a report from the prospective Lynch syndrome database. Gut, 66(9):1657–1664, 2017b.

Pål Møller, Toni T Seppälä, Inge Bernstein, Elke Holinski-Feder, Paulo Sala, D Gareth Evans, Annika Lindblom, Finlay Macrae, Ignacio Blanco, Rolf H Sijmons, et al. Cancer risk and survival in path_MMR carriers by gene and gender up to 75 years of age: a report from the Prospective Lynch Syndrome Database. Gut, 67(7):1306–1316, 2018.

Mohammad Movahedi, D Timothy Bishop, Finlay Macrae, Jukka-Pekka Mecklin, Gabriela Moeslein, Sylviane Olschwang, Diana Eccles, D Gareth Evans, Eamonn R Maher, Lucio Bertario, et al. Obesity, aspirin, and risk of colorectal cancer in carriers of hereditary colorectal cancer: a prospective investigation in the CAPP2 study. Journal of Clinical Oncology, 33(31):3591–3597, 2015.

Kasper Munch and Anders Krogh. Automatic generation of gene finders for eukaryotic species. BMC bioinformatics, 7(1):263, 2006.

Kevin Murphy, Yair Weiss, and Michael I Jordan. Loopy belief propagation for approximate inference: An empirical study. arXiv preprint arXiv:1301.6725, 2013.

Kanako Nakamura, Kouji Banno, Megumi Yanokura, Miho Iida, Masataka Adachi, Kenta Masuda, Arisa Ueki, Yusuke Kobayashi, Hiroyuki Nomura, Akira Hirasawa, et al. Features of ovarian cancer in Lynch syndrome. Molecular and clinical oncology, 2(6):909–916, 2014.

Joseph Irwin Naus. Clustering of Random Points in Line and Plane. Harvard University Press, 1963. URL `https://books.google.fr/books?id=uWOOGwAACAAJ`.

Wayne Nelson. Hazard plotting for incomplete failure data. Journal of Quality Technology, 1(1):27–52, 1969.

Wayne Nelson. Theory and applications of hazard plotting for censored failure data. Technometrics, 14(4):945–966, 1972.

Pierre Nicodeme, Bruno Salvy, and Philippe Flajolet. Motif statistics. Theoretical Computer Science, 287(2):593–617, 2002.

Dennis Nilsson. The computation of moments of decomposable functions in probabilistic expert systems. In Proceedings of the Third International Symposium on Adaptive Systems, pages 116–21, 2001.

Dennis Nilsson and Jacob Goldberger. Sequentially finding the N-best list in hidden Markov models. In International joint conference on artificial intelligence, volume 17, pages 1280–1285. LAWRENCE ERLBAUM ASSOCIATES LTD, 2001.

Yue S Niu and Heping Zhang. The screening and ranking algorithm to detect DNA copy number variations. The annals of applied statistics, 6(3):1306, 2012.

Jonathan A Nowak, Matthew B Yurgelun, Jacqueline L Bruce, Vanesa Rojas-Rudilla, Dimity L Hall, Priyanka Shivdasani, Elizabeth P Garcia, Agoston T Agoston, Amitabh Srivastava, Shuji Ogino, et al. Detection of mismatch repair deficiency and microsatellite instability in colorectal adenocarcinoma by targeted

next-generation sequencing. The Journal of Molecular Diagnostics, 19(1):84–91, 2017.

Grégory Nuel. LD-SPatt: large deviations statistics for patterns on Markov chains. Journal of Computational Biology, 11(6):1023–1033, 2004.

Grégory Nuel. Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. Journal of Applied Probability, 45(1):226–243, 2008.

Grégory Nuel. On the first k moments of the random count of a pattern in a multistate sequence generated by a Markov source. Journal of Applied Probability, 47(4): 1105–1123, 2010.

Grégory Nuel. Moments of the Count of a Regular Expression in a Heterogeneous Random Sequence. Methodology and Computing in Applied Probability, 21(3): 875–887, 2019.

Grégory Nuel, Leslie Regad, Juliette Martin, and Anne-Claude Camproux. Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data. Algorithms for Molecular Biology, 5(1): 1–18, 2010.

Grégory Nuel, Alexandra Lefebvre, and Olivier Bouaziz. Computing individual risks based on family history in genetic disease in the presence of competing risks. Computational and mathematical methods in medicine, 2017, 2017.

David Oakes. Direct calculation of the information matrix via the EM. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(2):479–482, 1999.

Jeffrey R O'Connell and Daniel E Weeks. The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set–recoding and fuzzy inheritance. Nature genetics, 11(4):402–408, 1995.

Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics, 5(4):557–572, 2004.

Jurg Ott. Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. American journal of human genetics, 26(5):588, 1974.

Jurg Ott. Analysis of human genetic linkage. JHU Press, 1999.

Michael J Overman, Ray McDermott, Joseph L Leach, Sara Lonardi, Heinz-Josef Lenz, Michael A Morse, Jayesh Desai, Andrew Hill, Michael Axelson, Rebecca A Moss, et al. Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. The Lancet Oncology, 18(9):1182–1191, 2017.

Michael J Overman, Sara Lonardi, Ka Yeung Mark Wong, Heinz-Josef Lenz, Fabio Gelsomino, Massimo Aglietta, Michael A Morse, Eric Van Cutsem, Ray McDermott, Andrew Hill, et al. Durable clinical benefit with nivolumab plus ipilimumab in DNA mismatch repair-deficient/microsatellite instability-high metastatic colorectal cancer. 2018.

Grier P Page, Varghese George, Rodney C Go, Patricia Z Page, and David B Allison. Are we there yet?: Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits, 2003.

Tuya Pal, Mohammad R Akbari, Philippe Sun, JH Lee, William J Fulp, Zachary J Thompson, D Coppola, Santo V Nicosia, Thomas A Sellers, John R McLaughlin, et al. Frequency of mutations in mismatch repair genes in a population-based study of women with ovarian cancer. British journal of cancer, 107(10):1783–1790, 2012.

Mala Pande, Patrick M Lynch, John L Hopper, Mark A Jenkins, Steve Gallinger, Robert W Haile, Loic LeMarchand, Noralane M Lindor, Peter T Campbell, Polly A Newcomb, et al. Smoking and colorectal cancer in Lynch syndrome: results from the Colon Cancer Family Registry and the University of Texas MD Anderson Cancer Center. Clinical cancer research, 16(4):1331–1339, 2010.

Michael T Parsons, Daniel D Buchanan, Bryony Thompson, Joanne P Young, and Amanda B Spurdle. Correlation of tumour BRAF mutations and MLH1 methylation with germline mismatch repair (MMR) gene mutation status: a literature review assessing utility of tumour features for MMR variant classification. Journal of medical genetics, 49(3):151–157, 2012.

Aniruddh P Patel, Minxian Wang, Akl C Fahed, Heather Mason-Suares, Deanna Brockman, Renee Pelletier, Sami Amr, Kalotina Machini, Megan Hawley, Leora Witkowski, et al. Association of rare pathogenic DNA variants for familial hypercholesterolemia, hereditary breast and ovarian cancer syndrome, and lynch syndrome with disease risk in adults according to family history. JAMA network open, 3(4):e203959–e203959, 2020.

Judea Pearl. Reverend Bayes on inference engines: A distributed hierarchical approach. Cognitive Systems Laboratory, School of Engineering and Applied Science ?, 1982.

Judea Pearl. Fusion, propagation, and structuring in belief networks. Artificial intelligence, 29(3):241–288, 1986.

Judea Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufman, San Mateo, 1988.

Danilo Pellin and Clelia Di Serio. A novel scan statistics approach for clustering identification and comparison in binary genomic data. BMC bioinformatics, 17 (11):61–71, 2016.

Nathalie Peyrard, M-J Cros, Simon de Givry, Alain Franc, Stephane Robin, Regis Sabbadin, Thomas Schiex, and Matthieu Vignes. Exact or approximate inference in graphical models: why the choice is dictated by the treewidth, and how variable elimination can be exploited. Australian & New Zealand Journal of Statistics, 61 (2):89–133, 2019.

Ruth M Pfeiffer, Yikyung Park, Aimée R Kreimer, James V Lacey Jr, David Pee, Robert T Greenlee, Saundra S Buys, Albert Hollenbeck, Bernard Rosner, Mitchell H Gail, et al. Risk prediction for breast, endometrial, and ovarian cancer in white women aged 50 y or older: derivation and validation from population-based cohort studies. PLoS Med, 10(7):e1001492, 2013.

Franck Picard, Stephane Robin, Marc Lavielle, Christian Vaisse, and Jean-Jacques Daudin. A statistical approach for array CGH data analysis. BMC bioinformatics, 6(1):27, 2005.

Bernard Prum. La démarche statistique. Cépaduès, 2010.

Bernard Prum, François Rodolphe, and Elisabeth De Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. Journal of the Royal Statistical Society: Series B (Methodological), 57(1):205–220, 1995.

Zhaohui S Qin, Jianjun Yu, Jincheng Shen, Christopher A Maher, Ming Hu, Shanker Kalyana-Sundaram, Jindan Yu, and Arul M Chinnaiyan. HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. BMC bioinformatics, 11(1):369, 2010.

Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286, 1989.

Antonio Raffone, Antonio Travaglino, Marco Cerbone, Annarita Gencarelli, Antonio Mollo, Luigi Insabato, and Fulvio Zullo. Diagnostic Accuracy of Immunohistochemistry for Mismatch Repair Proteins as Surrogate of Microsatellite Instability Molecular Testing in Endometrial Cancer. Pathology oncology research: POR, 2020.

Peter F Rambau, Máire A Duggan, Prafull Ghatage, Khadija Warfa, Helen Steed, Renee Perrier, Linda E Kelemen, and Martin Köbel. Significant frequency of MSH2/MSH6 abnormality in ovarian endometrioid carcinoma supports histotype-specific Lynch syndrome screening in ovarian carcinomas. Histopathology, 69(2): 288–297, 2016.

Mireille Régnier. A unified approach to word occurrence probabilities. Discrete applied mathematics, 104(1-3):259–280, 2000.

Gesine Reinert, Sophie Schbath, and Michael S Waterman. Probabilistic and statistical properties of words: an overview. Journal of Computational Biology, 7(1-2): 1–46, 2000.

Stéphane Robin and Jean-Jacques Daudin. Exact distribution of word occurrences in a random sequence of letters. Journal of Applied Probability, 36(1):179–193, 1999.

Stéphane Robin, Stéphane Robin, F Rodolphe, and S Schbath. DNA, words and models: statistics of exceptional words. Cambridge University Press, 2005.

Miguel A Rodriguez-Bigas, C Richard Boland, Stanley R Hamilton, Donald E Henson, Sudhir Srivastava, Jeremy R Jass, P Meera Khan, Henry Lynch, Thomas Smyrk, Manuel Perucho, et al. A National Cancer Institute workshop on hereditary nonpolyposis colorectal cancer syndrome: meeting highlights and Bethesda guidelines. Journal of the National Cancer Institute, 89(23):1758–1762, 1997.

María Rodríguez-Soler, Lucía Pérez-Carbonell, Carla Guarinos, Pedro Zapater, Adela Castillejo, Victor M Barberá, Miriam Juárez, Xavier Bessa, Rosa M Xicola, Juan Clofent, et al. Risk of cancer in cases of suspected lynch syndrome without germline mutation. Gastroenterology, 144(5):926–932, 2013.

Donald J Rose, R Endre Tarjan, and George S Lueker. Algorithmic aspects of vertex elimination on graphs. SIAM Journal on computing, 5(2):266–283, 1976.

Burkhard Rost, Piero Fariselli, and Rita Casadio. Topology prediction for helical transmembrane proteins at 86% accuracy–Topology prediction at 86% accuracy. Protein Science, 5(8):1704–1718, 1996.

Neil AJ Ryan, Raymond McMahon, Simon Tobi, Tristan Snowsill, Shona Esquibel, Andrew J Wallace, Sancha Bunstone, Naomi Bowers, Ioana E Mosneag, Sarah J Kitson, et al. The proportion of endometrial tumours associated with Lynch syndrome (PETALS): A prospective cross-sectional study. PLoS medicine, 17(9): e1003263, 2020.

Philippe Saint Pierre. Modèles multi-états de type Markovien et application à l'asthme. PhD thesis, Université Montpellier I, 2005.

Stephen J Salipante, Sheena M Scroggins, Heather L Hampel, Emily H Turner, and Colin C Pritchard. Microsatellite instability detection by next generation sequencing. Clinical chemistry, 60(9):1192–1199, 2014.

Daniel J Sargent, Silvia Marsoni, Genevieve Monges, Stephen N Thibodeau, Roberto Labianca, Stanley R Hamilton, Amy J French, Brian Kabat, Nathan R Foster, Valter Torri, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. Journal of Clinical Oncology, 28(20):3219, 2010.

Sophie Schbath. Compound Poisson approximation of word counts in DNA sequences. ESAIM: probability and statistics, 1:1–16, 1997.

Kathleen M Schmeler, Henry T Lynch, Lee-may Chen, Mark F Munsell, Pamela T Soliman, Mary Beth Clark, Molly S Daniels, Kristin G White, Stephanie G Boyd-Rogers, Peggy G Conrad, et al. Prophylactic surgery to reduce the risk of gynecologic cancers in the Lynch syndrome. New England Journal of Medicine, 354(3): 261–269, 2006.

Richard Schwartz and Y-L Chow. The N-best algorithms: an efficient and exact procedure for finding the N most likely sentence hypotheses. In International Conference on Acoustics, Speech, and Signal Processing, pages 81–84. IEEE, 1990.

Rishabh Sehgal, Kieran Sheahan, Patrick R O'Connell, Ann M Hanly, Sean T Martin, and Desmond C Winter. Lynch syndrome: an updated review. Genes, 5(3):497–507, 2014.

Leigha Senter, Mark Clendenning, Kaisa Sotamaa, Heather Hampel, Jane Green, John D Potter, Annika Lindblom, Kristina Lagerstedt, Stephen N Thibodeau, Noralane M Lindor, et al. The clinical phenotype of Lynch syndrome due to germ-line PMS2 mutations. Gastroenterology, 135(2):419–428, 2008.

Ross D Shachter. Probabilistic inference and influence diagrams. Operations research, 36(4):589–604, 1988.

Glenn R Shafer and Prakash P Shenoy. Probability propagation. Annals of mathematics and Artificial Intelligence, 2(1-4):327–351, 1990.

Mark Silberstein, Anna Tzemach, Nickolay Dovgolevsky, Maáyan Fishelson, Assaf Schuster, and Dan Geiger. Online system for faster multipoint linkage analysis via parallel execution on thousands of personal computers. The American Journal of Human Genetics, 78(6):922–935, 2006.

Emily A Sloan, Kari L Ring, Brian C Willis, Susan C Modesitt, and Anne M Mills. PD-L1 expression in mismatch repair-deficient endometrial carcinomas, including lynch syndrome-associated and MLH1 promoter hypermethylated tumors. The American journal of surgical pathology, 41(3):326–333, 2017.

Eric Sobel and Kenneth Lange. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. American journal of human genetics, 58(6):1323, 1996.

Eric Sobel, Haydar Sengul, and Daniel E Weeks. Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. Human heredity, 52(3):121–131, 2001.

Eric Sobel, Jeanette C Papp, and Kenneth Lange. Detection and integration of genotyping errors in statistical genetics. The American Journal of Human Genetics, 70(2):496–508, 2002.

Zbyslaw Sondka, Sally Bamford, Charlotte G Cole, Sari A Ward, Ian Dunham, and Simon A Forbes. The COSMIC Cancer Gene Census: describing genetic

dysfunction across all human cancers. Nature Reviews Cancer, 18(11):696–705, 2018.

Erik LL Sonnhammer, Gunnar Von Heijne, Anders Krogh, et al. A hidden Markov model for predicting transmembrane helices in protein sequences. In Ismb, volume 6, pages 175–182, 1998.

Isabelle Sourrouille, Florence Coulet, Jeremie H Lefevre, Chrystelle Colas, Mélanie Eyries, Magali Svrcek, Armelle Bardier-Dupas, Yann Parc, and Florent Soubrier. Somatic mosaicism and double somatic hits can lead to MSI colorectal tumors. Familial cancer, 12(1):27–33, 2013.

Zsofia K Stadler, Francesca Battaglin, Sumit Middha, Jaclyn F Hechtman, Christina Tran, Andrea Cercek, Rona Yaeger, Neil H Segal, Anna M Varghese, Diane L Reidy-Lagunes, et al. Reliable detection of mismatch repair deficiency in colorectal cancers using mutational load in next-generation sequencing panels. Journal of Clinical Oncology, 34(18):2141, 2016.

Ellen Stelloo, Remi A Nout, Elisabeth M Osse, Ina J Jürgenliemk-Schulz, Jan J Jobsen, Ludy C Lutgens, Elzbieta M van der Steen-Banasik, Hans W Nijman, Hein Putter, Tjalling Bosse, et al. Improved risk assessment by integrating molecular and clinicopathological factors in early-stage endometrial cancer?combined analysis of the PORTEC cohorts. Clinical cancer research, 22(16):4215–4224, 2016.

Ellen Stelloo, AML Jansen, Elisabeth M Osse, Remi A Nout, CL Creutzberg, D Ruano, DN Church, H Morreau, VTHBM Smit, T van Wezel, et al. Practical guidance for mismatch repair-deficiency testing in endometrial cancer. Annals of Oncology, 28(1):96–102, 2017.

Nirosha Suraweera, Alex Duval, Maryline Reperant, Christelle Vaury, Daniela Furlan, Karen Leroy, Raquel Seruca, Barry Iacopetta, and Richard Hamelin. Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR. Gastroenterology, 123(6):1804–1811, 2002.

Walter S Sutton. On the morphology of the chromoso group in Brachystola magna. The Biological Bulletin, 4(1):24–39, 1902.

Walter S Sutton. The chromosomes in heredity. The Biological Bulletin, 4(5):231–250, 1903.

Magali Svrcek, Olivier Lascols, Romain Cohen, Ada Collura, Vincent Jonchère, Jean-François Fléjou, Olivier Buhard, and Alex Duval. MSI/MMR-deficient tumor diagnosis: Which standard for screening and for diagnosis? Diagnostic modalities for the colon and other sites: Differences between tumors. Bulletin du cancer, 106 (2):119–128, 2019.

Sapna Syngal, Randall E Brand, James M Church, Francis M Giardiello, Heather L Hampel, and Randall W Burt. ACG clinical guideline: genetic testing and management of hereditary gastrointestinal cancer syndromes. The American journal of gastroenterology, 110(2):223, 2015.

Daiya Takai and Peter A Jones. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proceedings of the national academy of sciences, 99(6): 3740–3745, 2002.

Aline Talhouk, Melissa K McConechy, Samuel Leung, Hector H Li-Chang, JS Kwon, N Melnyk, W Yang, J Senz, N Boyd, AN Karnezis, et al. A clinically applicable molecular-based classification for endometrial cancers. British journal of cancer, 113(2):299–310, 2015.

Christoph Tamm, Herman S Shapiro, Rakoma Lipshitz, and Erwin Chargaff. Distribution density of nucleotides within a desoxyribonucleic acid chain. Journal of Biological Chemistry, 203(2):673–688, 1953.

Robert E Tarjan and Mihalis Yannakakis. Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. SIAM Journal on computing, 13(3):566–579, 1984.

Sanne W Ten Broeke, Heleen M van der Klift, Carli MJ Tops, Stefan Aretz, Inge Bernstein, Daniel D Buchanan, Albert de la Chapelle, Gabriel Capella, Mark Clendenning, Christoph Engel, et al. Cancer risks for PMS2-associated Lynch syndrome. Journal of Clinical Oncology, 36(29):2961, 2018.

Duncan C Thomas et al. Statistical methods in genetic epidemiology. Oxford University Press, 2004.

Bryony A Thompson, David E Goldgar, Carol Paterson, Mark Clendenning, Rhiannon Walters, Sven Arnold, Michael T Parsons, Walsh Michael D, Steven Gallinger, Robert W Haile, et al. A multifactorial likelihood model for MMR gene variant classification incorporating probabilities based on sequence bioinformatics and tumor characteristics: a report from the Colon Cancer Family Registry. Human mutation, 34(1):200–209, 2013.

Bryony A Thompson, Amanda B Spurdle, John-Paul Plazzer, Marc S Greenblatt, Kiwamu Akagi, Fahd Al-Mulla, Bharati Bapat, Inge Bernstein, Gabriel Capella, Johan T den Dunnen, Desiree du Sart, Aurelie Fabre, Michael P Farrell, Susan M Farrington, Ian M Frayling, Thierry Frebourg, David E Goldgar, Christopher D Heinen, Elke Holinski-Feder, Maija Kohonen-Corish, Kristina Lagerstedt Robinson, Suet Yi Leung, Alexandra Martins, Pal Moller, Monika Morak, Minna Nystrom, Paivi Peltomaki, Marta Pineda, Ming Qi, Rajkumar Ramesar, Lene Juel Rasmussen, Brigitte Royer-Pokora, Rodney J Scott, Rolf Sijmons, Sean V Tavtigian, Carli M Tops, Thomas Weber, Juul Wijnen, Michael O Woods, Finlay Macrae, Genuardi, on behalf of InSiGHT, and The InSiGHT collaborators. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nature Genetics, 46 (2):107–115, 2014.

Elizabeth A Thompson. Monte Carlo likelihood in genetic mapping. Statistical Science, pages 355–366, 1994.

Elizabeth A Thompson. Statistical inference from genetic data on pedigrees. In NSF-CBMS regional conference series in probability and statistics, pages i–169. JSTOR, 2000.

Elizabeth A Thompson and Simon C Heath. Estimation of conditional multilocus gene identity among relatives. Lecture Notes-Monograph Series, pages 95–113, 1999.

Michalis K Titsias, Christopher C Holmes, and Christopher Yau. Statistical inference in hidden markov models using k-segment constraints. Journal of the American Statistical Association, 111(513):200–215, 2016.

Liviu R Totir, Rohan L Fernando, and Joseph Abraham. An efficient algorithm to compute marginal posterior genotype probabilities for every member of a pedigree with loops. Genetics Selection Evolution, 41(1):52, 2009.

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. Signal Processing, 167:107299, 2020.

Jonathan Tyrer, Stephen W Duffy, and Jack Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. Statistics in medicine, 23 (7):1111–1130, 2004.

Asad Umar, C Richard Boland, Jonathan P Terdiman, Sapna Syngal, Albert de la Chapelle, Josef Rüschoff, Richard Fishel, Noralane M Lindor, Lawrence J Burgart, Richard Hamelin, et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. Journal of the National Cancer Institute, 96(4):261–268, 2004.

Margot GF van Lier, Celine HM Leenen, Anja Wagner, Dewkoemar Ramsoekh, Hendrikus J Dubbink, Ans MW van den Ouweland, Pieter J Westenend, Eelco JR de Graaf, Leonieke MM Wolters, Wietske W Vrijland, et al. Yield of routine molecular analyses in colorectal cancer patients? 70 years to detect underlying Lynch syndrome. The Journal of pathology, 226(5):764–774, 2012.

Hans F Vasen, A Stormorken, FH Menko, FM Nagengast, JH Kleibeuker, G Griffioen, BG Taal, P Moller, and JT Wijnen. MSH2 mutation carriers are at higher risk of cancer than MLH1 mutation carriers: a study of hereditary nonpolyposis colorectal cancer families. 2001.

Hans FA Vasen, J-P Mecklin, P Meera Khan, and Henry T Lynch. The international collaborative group on hereditary non-polyposis colorectal cancer (ICG-HNPCC). Diseases of the Colon & Rectum, 34(5):424–425, 1991.

Hans FA Vasen, Patrice Watson, Jukka-Pekka Mecklin, Henry T Lynch, et al. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. Gastroenterology, 116(6):1453–1456, 1999.

Hans FA Vasen, Ignacio Blanco, Katja Aktan-Collan, Jessica P Gopie, Angel Alonso, Stefan Aretz, Inge Bernstein, Lucio Bertario, John Burn, Gabriel Capella, et al. Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts. Gut, 62(6):812–823, 2013.

Jean-Philippe Vert and Kevin Bleakley. Fast detection of multiple change-points shared by many signals using group LARS. Advances in neural information processing systems, 23:2343–2351, 2010.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE transactions on Information Theory, 13(2): 260–269, 1967.

Gunnar Von Heijne. Membrane protein structure prediction: hydrophobicity analysis and the positive-inside rule. Journal of molecular biology, 225(2):487–494, 1992.

Janet R Vos, Ingrid E Fakkert, Liesbeth Spruijt, Riki W Willems, Sera Langenveld, Arjen R Mensenkamp, Edward M Leter, Iris D Nagtegaal, Marjolijn JL Ligtenberg, Nicoline Hoogerbrugge, et al. Evaluation of yield and experiences of age-related molecular investigation for heritable and nonheritable causes of mismatch repair deficient colorectal cancer to identify Lynch syndrome. International journal of cancer, 147(8):2150–2158, 2020.

Martin J Wainwright and Michael Irwin Jordan. Graphical models, exponential families, and variational inference. Now Publishers Inc, 2008.

Robyn L Ward, Timothy Dobbins, Noralane M Lindor, Robert W Rapkins, and Megan P Hitchins. Identification of constitutional MLH1 epimutations and promoter variants in colorectal cancer patients from the Colon Cancer Family Registry. Genetics in medicine, 15(1):25–35, 2013.

Alfred Scott Warthin. Heredity with reference to carcinoma: as shown by the study of the cases examined in the pathological laboratory of the University of Michigan, 1895-1913. Archives of internal medicine, 12(5):546–555, 1913.

James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature, 171(4356):737–738, 1953.

Scott M Weissman, Randall Burt, James Church, Steve Erdman, Heather Hampel, Spring Holter, Kory Jasperson, Matt F Kalady, Joy Larsen Haidle, Henry T Lynch, et al. Identification of individuals at risk for Lynch syndrome using targeted evaluations and genetic testing: National Society of Genetic Counselors and the Collaborative Group of the Americas on Inherited Colorectal Cancer joint practice guideline. Journal of genetic counseling, 21(4):484–493, 2012.

Andreas Wienke. Frailty models in survival analysis. CRC press, 2010.

Ellen M Wijsman, Joseph H Rothstein, and Elizabeth A Thompson. Multipoint linkage analysis with many multiallelic or dense diallelic markers: Markov Chain–Monte Carlo provides practical approaches for genome scans on general Pedigrees. The American Journal of Human Genetics, 79(5):846–858, 2006.

Aung Ko Win, Robert J MacInnis, James G Dowty, and Mark A Jenkins. Criteria and prediction models for mismatch repair gene mutations: a review. Journal of medical genetics, 50(12):785–793, 2013.

Aung Ko Win, Mark A Jenkins, James G Dowty, Antonis C Antoniou, Andrew Lee, Graham G Giles, Daniel D Buchanan, Mark Clendenning, Christophe Rosty, Dennis J Ahnen, et al. Prevalence and penetrance of major genes and polygenes for colorectal cancer. Cancer Epidemiology and Prevention Biomarkers, 26(3): 404–412, 2017.

Sewall Wright. Correlation and Causation. Journal of Agricultural Research, 20: 557–585, 1921.

Sewall Wright. The method of path coefficients. The annals of mathematical statistics, 5(3):161–215, 1934.

Ying Wu, Maran JW Berends, Rob GJ Mensink, Claudia Kempinga, Rolf H Sijmons, Ate GJ van der Zee, Harry Hollema, Jan H Kleibeuker, Charles HCM Buys, and Robert MW Hofstra. Association of hereditary nonpolyposis colorectal cancer–related tumors displaying low microsatellite instability with MSH6 germline mutations. The American Journal of Human Genetics, 65(5):1291–1298, 1999.

Mihalis Yannakakis. Computing the minimum fill-in is NP-complete. SIAM Journal on Algebraic Discrete Methods, 2(1):77–79, 1981.

Byung-Jun Yoon. Hidden Markov models and their applications in biological sequence analysis. Current genomics, 10(6):402–415, 2009.

Rianon Zaman, Shahana Yasmin Chowdhury, Mahmood A Rashid, Alok Sharma, Abdollah Dehzangi, and Swakkhar Shatabda. Hmmbinder: DNA-binding protein prediction using HMM profile based features. BioMed research international, 2017, 2017.

Nancy R Zhang, Benjamin Yakir, Li C Xia, and David Siegmund. Scan statistics on Poisson random fields with applications in genomics. The Annals of Applied Statistics, 10(2):726–755, 2016.

Bo Zhao and Joseph Glaz. Scan statistics for detecting a local change in variance for normal data with unknown population variance. Statistics & Probability Letters, 110:137–145, 2016.

Bo Zhao and Joseph Glaz. Scan statistics for detecting a local change in variance for two-dimensional normal data. Communications in Statistics-Theory and Methods, 46(11):5517–5530, 2017.

Gang Zhao and Erwin London. An amino acid transmembrane tendency scale that approaches the theoretical limit to accuracy for prediction of transmembrane helices: relationship to biological hydrophobicity. Protein science, 15(8):1987–2001, 2006.