Évaluation de modèles pronostiques issus de l'analyse dutranscriptome

Caroline Truntzer

N^o d'ordre 78-2007 Année 2007

THESE

présentée

devant l'UNIVERSITE CLAUDE BERNARD - LYON 1

pour l'obtention

du DIPLOME DE DOCTORAT

(arrêté du 25 avril 2002)

présentée et soutenue publiquement le 8 juin 2007

par

Caroline TRUNTZER

Titre

Évaluation de modèles pronostiques issus de l'analyse du transcriptome

Directeur de thèse : Pr. Pascal ROY

JURY :	Pr. Philippe BESSE	Rapporteur
	Pr. Jean-Jacques DAUDIN	Rapporteur
	Pr. Christian GAUTIER	Examinateur
	Pr. Dominique MOUCHIROUD	Président
	Pr. Pascal ROY	Examinateur

UNIVERSITE CLAUDE BERNARD - LYON I

Président de l'Université	**M. le Professeur L. COLLET**
Vice-Président du Conseil Scientifique	M. le Professeur J.F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur R. GARRONE
Vice-Présidente du Conseil des Etudes et de la Vie Universitaire	M. le Professeur G. ANNAT
Secrétaire Général	M. M. GIRARD

SECTEUR SANTE

Composantes

UFR de Médecine Lyon R.T.H. Laënnec	Directeur : M. le Professeur D. VITAL-DURAND
UFR de Médecine Lyon Grange-Blanche	Directeur : M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	Directeur : M. le Professeur F. MAUGUIERE
UFR de Médecine Lyon-Sud	Directeur : M. le Professeur F.N. GILLY
UFR d'Odontologie	Directeur : M. O. ROBIN
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. LOCHER
Institut Techniques de Réadaptation	Directeur : M. le Professeur L. COLLET
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

SECTEUR SCIENCES

Composantes

UFR de Physique	Directeur : M. le Professeur A. HOAREAU
UFR de Biologie	Directeur : M. le Professeur H. PINON
UFR de Mécanique	Directeur : M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	Directeur : M. le Professeur A. BRIGUET
UFR Sciences de la Terre	Directeur : M. le Professeur P. HANTZPERGUE
UFR de Mathématiques	Directeur : M. le Professeur M. CHAMARIE
UFR d'Informatique	Directeur : M. le Professeur M. EGEA
UFR de Chimie Biochimie	Directeur : M. le Professeur J.P. SCHARFF
UFR STAPS	Directeur : M. le Professeur R. MASSARELLI
Observatoire de Lyon	Directeur : M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. LIETO
IUT A	Directeur : M. le Professeur M. C. COULET
IUT B	Directeur : M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J.C. AUGROS

Remerciements

Je remercie tout d'abord la Ligue Nationale Contre le Cancer dont le soutien financier m'a permis d'effectuer ma thèse dans de bonnes conditions.

Je remercie les rapporteurs de cette thèse, Philippe Besse et Jean-Jacques Daudin, ainsi que Dominique Mouchiroud pour l'intérêt qu'ils ont porté à mon travail.

J'aimerais ensuite remercier Pascal Roy et René Ecochard de m'avoir accueillie au sein de leur équipe.

Plus particulièrement, je remercie Pascal Roy de m'avoir soutenue pendant ces trois années et de m'avoir ouvert des perspectives dans le monde des biostatistiques.

Je remercie également Christian Gautier pour sa disponibilité et ses conseils avisés.

Je remercie aussi chaleureusement tous les membres de l'équipe Biostatistique-Santé, avec une pensée particulière pour Maud, pour leur soutien scientifique et moral, et pour les discussions diverses que nous avons pu avoir, que ce soit autour d'un article scientifique...ou d'une part de gâteau...ça a été un plaisir de travailler parmi vous.

Un grand merci à mes amis, et plus particulièrement Catherine, Céline, Claire, Isabelle et Nadège, qui ont su m'écouter dans les moments difficiles, se réjouir avec moi dans les bons moments, et tout simplement me faire profiter de leur amitié.

Je termine enfin par remercier ma famille, mes parents et mon frère pour leur soutien constant. Je leur dédie mon mémoire de thèse.

Table des figures

Liste des tableaux

Table des matières

Première partie

Introduction

Le 13 décembre 2006 paraît dans le quotidien *Le Monde* un article intitulé "Mieux identifier les cancers pour mieux les traiter" [1]. Cet article présente une étude, menée pour la première fois en France, qui doit évaluer l'intérêt de l'utilisation des biopuces pour le choix de la chimiothérapie la plus adaptée à la patiente pour des femmes atteintes d'un cancer du sein. On peut y lire : "Pour décider du traitement à engager, les médecins se fondent sur des critères relatifs à l'importance de la prolifération tumorale dans le tissu, l'envahissement ou non des ganglions par les cellules malignes, [...]. Avec le développement des puces à ADN, les praticiens espèrent être capables d'établir un meilleur pronostic d'évolution des cancers [...] La question est de savoir, pour prédire l'évolution de la tumeur, si le profil génomique sera un facteur de pronostic plus puissant que le nombre de ganglions envahis."

L'auteur pose une question essentielle, à une époque où les analyses transcriptomiques prennent une place de plus en plus prédominante en recherche clinique, et où d'autres biotechnologies modernes regroupées sous le terme "omique", comme la protéomique par exemple, voient le jour. Pour une seule étude, ces technologies génèrent une quantité d'information sans commune mesure avec les données disponibles jusqu'alors. Ces études doivent conduire à l'identification de nouveaux biomarqueurs pour améliorer la prédiction du pronostic des patients.

Pour répondre à la question posée par l'article, il est nécessaire de répondre aux nombreuses questions méthodologiques posées par l'analyse simultanée d'un si grand nombre de variables. Cette démarche revêt au moins deux aspects : une validation des méthodes de prédiction employées grâce à une meilleure compréhension de leurs propriétés dans ce nouveau contexte d'utilisation, et une évaluation des capacités prédictives réelles fournie par l'analyse du transcriptome.

Un nombre considérable de méthodes de prédiction a été proposé pour l'analyse des biopuces. Face à cette multitude de méthodes, il n'existe pas de consensus préconisant une méthode standard adaptée quel que soit les données disponibles pour l'étude. Il semble aujourd'hui plus utile de mieux comprendre les méthodes disponibles que de continuer à en développer de nouvelles. En effet, la validation des biomarqueurs, indispensable pour leur utilisation clinique en routine, ne peut pas se faire sans une validation de la méthodologie employée pour leur détection.

Alors que cette étape semble indispensable, ce travail est encore peu engagé dans la littérature actuelle. Nous avons choisi de contribuer à cette démarche par l'évaluation de l'influence de la structure des données sur les qualités prédictives de trois variantes d'analyse discriminante.

Ceci a fait l'objet de la première partie de mon travail, et propose une réponse au premier aspect soulevé par l'article. Le second aspect pose la question de la validité statistique des biomarqueurs identifiés, et plus précisément la question de l'évaluation de leur pouvoir prédictif réel.

Si la plupart des biomarqueurs clinico-biologiques classiques ont été identifiés et validés, les biomarqueurs transcriptomiques posent simultanément la question de leur sélection et de leur validation. Les deux types de biomarqueurs sont donc à des niveaux différents et il est nécessaire d'évaluer l'impact de la phase de sélection dont sont encore sujets les biomarqueurs transcriptomiques. C'est une étape importante préliminaire à la construction de modèles intégrant les deux types de biomarqueurs. Cette question a fait l'objet de la seconde partie de mon travail.

Ce manuscrit est organisé de la façon suivante. Dans une première partie sera présenté le contexte sous-jacent à l'analyse du transcriptome, aussi bien sous l'aspect biologique que méthodologique. Dans la deuxième partie sera décrit le travail qui a été conduit pour répondre aux deux aspects énoncés ci-dessus. Enfin, les perspectives engagées par ce travail seront exposées, en particulier l'ouverture à l'analyse du protéome.

Deuxième partie

Présentation du contexte

Chapitre 1

Contexte biologique

Un des objectifs majeurs de la recherche clinique est l'identification de nouveaux biomarqueurs. Dans le cas des cancers, on parle de marqueurs tumoraux ; ce sont des molécules, souvent des protéines ou des polypeptides, synthétisées par les cellules cancéreuses et présentes soit dans la tumeur, soit dans le sang ou les urines en quantités mesurables, et qui sont des indicateurs de l'état pathologique du patient.

Ces marqueurs peuvent être à usage diagnostique ou pronostique. Dans le premier cas, ils permettent de déterminer la maladie dont le patient est atteint. Dans le second cas, ils permettent de prédire, après le diagnostic, le degré de gravité et l'évolution ultérieure de cette maladie, y compris son issue. La connaissance du pronostic des patients atteints d'une maladie grave est l'un des éléments autorisant une prise en charge optimale de ceux-ci. Les résultats des études pronostiques permettent en effet d'identifier les patients de bon et mauvais pronostic, de proposer aux patients une prise en charge thérapeutique adaptée à leur pronostic, réservant les traitements les plus lourds aux patients de mauvais pronostic, et d'identifier des sous-groupes de patients candidats pouvant être inclus dans des essais thérapeutiques. A l'heure actuelle, les biomarqueurs utilisés en routine dans la pratique clinique sont de nature clinico-biologique : le PSA (Prostate Specific Antigen) pour le cancer de la prostate, l'ACE (Antigène Carcino-Embryonnaire) pour les tumeurs gastro-intestinales, ou encore l'antigène CA 15-3 pour le cancer du sein, etc.

Depuis les années 90, la recherche de biomarqueurs a pris une nouvelle orientation, en s'intéressant aux produits de l'expression des gènes de cellules malades en comparaison de ceux de cellules normales. C'est ce nouveau type d'analyse, nommé analyse du transcriptome, qui est aujourd'hui au coeur des activités cliniques et de ce travail.

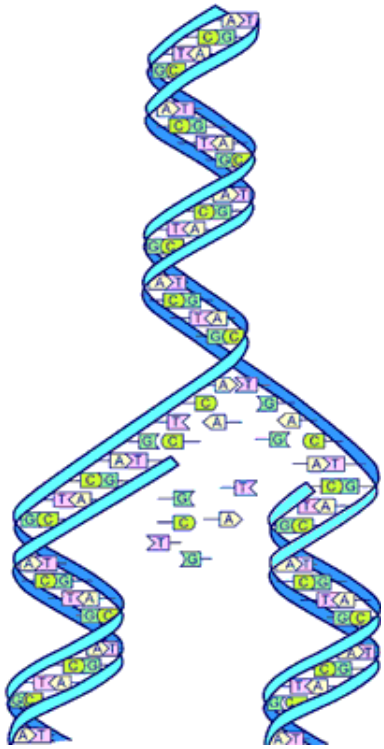


Fig. 1.1: *Structure de l'ADN*

1.1 Rappels de biologie moléculaire

La cellule est l'unité structurale, fonctionnelle et reproductrice constituant un être vivant. Dans chaque cellule d'un être humain se trouve le noyau, qui contient l'information génétique codée sous forme d'Acide DésoxyriboNucléique, ou ADN. L'ADN est constitué de deux brins en forme de double hélice. Un brin d'ADN est formé d'un enchaînement d'entités appelées nucléotides, un nucléotide étant formé d'un sucre, d'un groupement phosphate et d'une base azotée. Quatre bases azotées différentes peuvent être utilisées, ce qui conduit à quatre nucléotides différents : l'adénine (A), la cytosine (C), la guanine (G), et la thymine (T). Les bases sont complémentaires deux à deux : l'adénine s'associe avec la thymine et la guanine avec la cytosine. Le second brin d'ADN est complémentaire au premier, comme l'illustre la figure 1.1.

Sur l'ADN, le gène est l'unité de base de l'information génétique. Caractérisé par sa séquence de nucléotides, un gène permet la synthèse d'une protéine, caractérisée par sa séquence en acides aminés. Ces protéines serviront à leur tour à la construction et au bon fonctionnement de l'organisme vivant. Il peut s'agir de protéines servant à fabriquer des tissus, d'enzymes qui assureront par exemple la digestion des aliments, etc. L'homme possède entre 20 000 et 25 000 gènes, ce qui ne représente que 5% de son ADN : les gènes ne constituent qu'une partie du

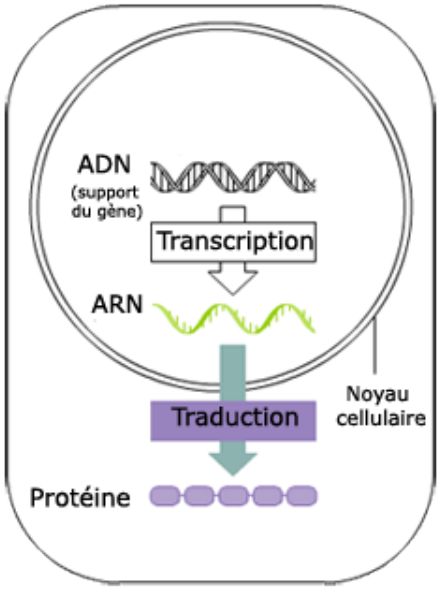


FIG. 1.2: *Schéma de la synthèse des protéines*

génome, celui-ci étant défini comme l'ensemble du matériel génétique d'un individu ou d'une espèce.

Le passage du gène à la protéine s'effectue en deux étapes, comme indiqué sur la figure 1.2 :

1. Transcription : la cellule recopie une partie de son ADN sous forme d'Acide RiboNucléique messager (ARNm). L'ensemble des ARNm issus de l'expression d'une partie du génome d'un tissu cellulaire ou d'un type de cellule constitue le transcriptome.
2. Traduction : alors que l'ADN ne peut sortir du noyau, l'ARNm est véhiculé hors du noyau et traduit en acides aminés, qui forment les protéines. L'ensemble des protéines présentes dans des conditions et à un moment donnés, constitue le protéome.

Si le génome est identique dans chacune des cellules d'un organisme donné, les gènes peuvent avoir une expression spécifique différenciée dans le temps (propre à un stade du développement), dans l'espace (propre à un type cellulaire, tissulaire ou organique) et/ou caractéristique d'un état donné (normal, pathologique ou en réponse à un stimulus particulier). Le mécanisme de transcription est hautement régulé. L'étude du transcriptome consiste à caractériser et à quantifier dans un tissu, dans un état et à un moment donné du développement, le niveau d'expression de gènes d'intérêt. Les biopuces sont un outil permettant cette quantification [2].

1.2 Principe des biopuces

1.2.1 Principe général

La technologie des biopuces est basée sur les propriétés d'hybridation[1] des nucléotides constituant les gènes. Partant de cette propriété, on greffe sur un support des oligonucléotides[2], appelés sondes, dont le rôle est de détecter par hybridation des cibles marquées complémentaires présentes dans le mélange à analyser. Le terme de "cible" ("target" en anglais) désigne l'ARNm que l'on cherche à identifier ou à quantifier, tandis que le terme de "sonde" ("probes" en anglais) désigne les molécules utilisées pour réaliser la détection, qui sont soit greffées sur le support, soit synthétisées in situ. Chaque sonde correspond à une séquence nucléotidique connue ; elle est présente en un nombre variable d'exemplaires, et possède une adresse connue sur le support. Les signaux d'hybridation correspondant à chaque sonde (sur la puce, on parle de spot) sont détectés selon le type de marquage par mesure radiographique ou par fluorescence, puis quantifiés. On fait l'hypothèse que ce signal est proportionnel à la quantité d'ARNm présent dans la cellule dont il provient. Les biopuces peuvent être élaborées suivant différentes technologies. Nous présenterons plus particulièrement deux des technologies les plus fréquemment utilisées : les puces à fluorescence sur lame de verre, et les puces à oligonucléotides de type Affymetrix.

1.2.1.1 Puces à fluorescence sur lames de verre

Dans ce cas, les sondes sont des oligonucléotides synthétisés de manière indépendantes, puis fixés sur des lames de verre. Les cibles sont extraites des échantillons biologiques à étudier et correspondent à des ARNm obtenus pour deux conditions d'intérêt, par exemple deux sous-types de tumeur. Par un mécanisme de transcription inverse[3], on obtient l'ADN complémentaire de l'ARN à étudier, auquel sont ajoutés des fluorochromes différents pour chacune des conditions. Les fluorochromes les plus fréquemment utilisés sont les cyanines Cy3 (couleur verte) et Cy5 (couleur rouge). Souvent, une condition de référence est utilisée, à laquelle sont comparées les autres conditions. Par des techniques d'analyse d'image, on obtient par ce procédé une image en fausses couleurs composée de spots allant du vert (cas où seul l'ADN des cellules de la condition 1 s'est fixé à la sonde), au rouge (cas où seul l'ADN des cellules de la condition 2 s'est fixé à la

[1] Association de chaînes d'acides nucléiques simple brin pour former des doubles brins, basée sur la complémentarité des séquences de nucléotides.

[2] Segment d'ADN simple brin composé de quelques dizaines de nucléotides.

[3] Ensemble des mécanismes moléculaires conduisant à la synthèse d'un ADN à partir de l'ARN.

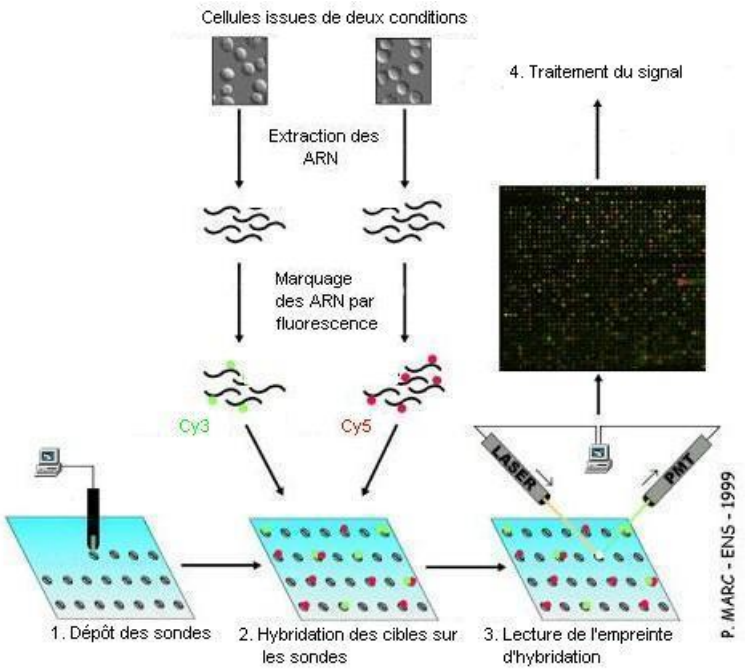


FIG. 1.3: *Principe des puces à fluorescence sur lame de verre*

sonde), en passant par le jaune (cas où l'ADN des deux types de cellules s'est fixé en quantité équivalente). C'est le rapport des intensités des fluorescences rouge sur verte de chaque spot qui est conservé pour les analyses ultérieures. On parle d'intensité relative.

Ce type de technologie permet d'atteindre une densité de 1 000 à 10 000 spots/cm^2. La figure 1.3 résume le principe de fonctionnement des puces à fluorescence.

1.2.1.2 Puces à oligonucléotides de type Affymetrix

Cette fois, les sondes sont des oligonucléotides de petite taille (25 mers[1]) qui sont synthétisés in situ. Chaque gène est représenté par un ensemble de deux types d'oligonucléotides appariés : 11 oligonucléotides complémentaires à la séquence du gène, qualifiés de "perfect-match" (PM), et 11 oligonucléotides différents du PM par une mutation du nucléotide situé en position centrale, et qualifiés de "mis-match" (MM). L'objectif de ces doublons est de contrôler les hybridations non spécifiques. L'ensemble des 22 oligonucléotides est qualifié de "probeset". Un seul type de fluorochrome est utilisé pour la cible : de la streptavidine couplée à la phycorythrine. Comme précédemment, un logiciel de traitement d'images génère une image de synthèse où chaque spot est représenté par des pixels dont l'intensité est fonction de la quantité d'ARN retenue par chaque séquence d'oligonucleotides. Pour chaque probeset, c'est la différence entre

[1]Enchaînement de 25 nucléotides.

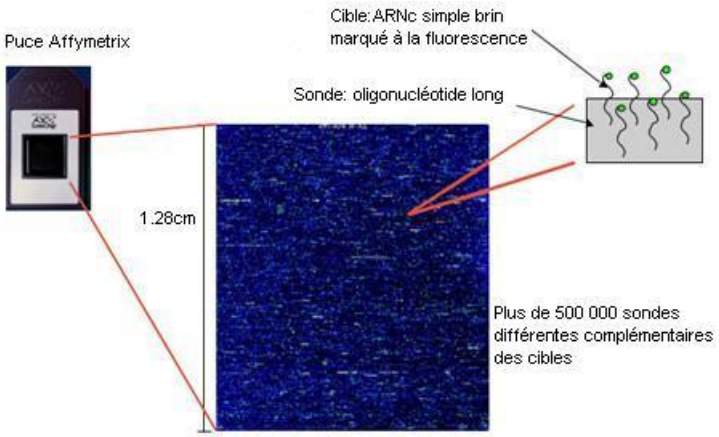


FIG. 1.4: *Principe des puces à oligonucéotides*

les PM et les MM qui est conservée pour les analyses ultérieures. On parle cette fois d'intensité absolue. Cette technologie permet d'atteindre une densité de 250 000 spots $/cm^2$. La figure 1.4 illustre le principe de fonctionnement de ce type de biopuces.

1.2.2 Étapes de l'analyse des biopuces

L'analyse des biopuces peut se décomposer en plusieurs étapes :

1.2.2.1 Plan expérimental

Il doit permettre de prendre en compte la question biologique posée ainsi que les contraintes expérimentales et matérielles. Il permet de définir à priori la taille des échantillons nécessaires à une analyse de qualité, le choix des conditions expérimentales. Il est nécessaire d'introduire des réplications dans l'analyse pour contrôler les sources de variabilités, qui sont de nature technique ou biologique.

1.2.2.2 Pré-traitement des données

Si les biopuces fournissent simultanément un grand nombre d'informations, de nombreuses sources de variabilités ou d'erreurs sont introduites à chaque étape de l'expérience biologique. Par exemple, les propriétés des fluorochromes sont différentes, ou les conditions de température et d'humidité, qui influent grandement l'étape d'hybridation, peuvent varier. On peut également rencontrer des problèmes de saturation des logiciels d'acquisition, etc. Toutes ces sources d'erreurs expérimentales introduisent une variabilité technique qui masque la variabilité biologique. Le rôle des étapes de pré-traitement est d'extraire cette variabilité biologique.

- Extraction du signal et analyse de l'image. Dans un premier temps, les spots doivent être localisés sur la puce. La segmentation permet ensuite de déterminer quelles régions de l'image correspondent à du signal, en le séparant du bruit de fond généré par des hybridations non spécifiques. Une fois le signal identifié, une méthode doit être choisie pour résumer le niveau d'expression associé à chaque spot (moyenne, médiane). Dans le cas des puces à oligonucléotides, une méthode supplémentaire doit être choisie pour résumer les niveaux d'expression des "probeset" en une seule mesure.
- Contrôles visuels de qualité. Des représentations graphiques simples permettent de repérer visuellement divers artefacts expérimentaux : boîtes à moustaches pour représenter les intensités de chaque biopuce, graphes "MA" pour les puces à fluorescence qui tracent le log-ratio des intensités (M) en fonction de la moyenne des log-intensités (A), etc. Ils donnent une première idée de l'importance des corrections à apporter.
- Soustraction du bruit de fond. Cette étape a pour but de soustraire au signal les pixels non associés aux spots. Elle est cependant controversée et il semble actuellement préférable de s'en affranchir.
- Normalisation. L'objectif est de corriger les données pour minimiser la part de variabilité non-biologique, et être à même de comparer plusieurs puces utilisant le même ensemble de gènes. Elle se fait à deux niveaux :
  - Intra-puces. Dans le cas des puces à fluorescences, différents effets sont corrigés par cette normalisation : effets des fluorochromes, effets spatiaux, effets d'aiguille...L'objectif de cette étape est de mettre tous les gènes d'une puce au même niveau.
  - Inter-puces. L'objectif de cette étape est de pouvoir comparer différentes puces. Cette fois, l'effet dont on cherche à s'affranchir est un "effet lame".

1.2.2.3 Analyse statistique des données

En cancérologie, les biopuces sont essentiellement utilisées pour la caractérisation de tumeurs. C'est la question biologique qui définit le type d'analyse à effectuer.

- Identifier des gènes marqueurs caractéristiques d'une classe de tumeurs : on utilisera pour cela des méthodes d'analyse différentielle.
- Identifier de nouvelles classes de tumeurs en utilisant le profil d'expression des gènes sans ajout de connaissance à priori : on peut également chercher à détecter des sous groupes de gènes, pour détecter par exemple de nouvelles voies métaboliques impliquées dans

un cancer particulier. On utilisera des méthodes d'analyse non supervisée (méthodes de classification). Ce sont essentiellement des techniques exploratoires en vue d'analyses plus poussées.

– Classer une tumeur parmi des classes connues : pour cela, on discrimine des classes connues à priori en vue de prédire le diagnostic ou le pronostic de nouveaux patients. On utilisera des méthodes d'analyse supervisée (méthodes de classement).

1.2.2.4 Validation et interprétation des résultats

Cette étape comprend la comparaison des résultats obtenus avec ceux d'autres plateformes, ou l'utilisation de jeux de données indépendants pour la validation des résultats. L'interprétation des résultats ne peut se faire sans un travail commun avec le clinicien investigateur.

Simon *et al.* [3], Miller *et al.* [4], puis Allison *et al.* [5] ont proposé des revues détaillées des problématiques liées à chaque point clé des études de biopuces, depuis le plan expérimental jusqu'à l'analyse statistique.

1.3 Exemples d'application des biopuces

L'utilisation de la technologie des biopuces dans le domaine médical a connu un développement exponentiel depuis son apparition. En particulier, c'est un outil de choix pour la comparaison entre le métabolisme de cellules "normales" et de cellules tumorales, question centrale dans la compréhension des mécanismes de genèse tumorale. Selon certaines études, l'analyse des biopuces permettrait d'identifier des sous-types non identifiables par les marqueurs cliniques ou histologiques classiques. Nous citerons ici quelques applications cliniques des biopuces, qui comptent parmi les plus fréquemment citées.

En 1999, Golub *et al.* [6] se basent sur l'expression des gènes pour le classement de patients atteints de lymphome en deux sous-types particuliers : leucémie aiguë lymphoblastique[1] (ALL), et leucémie aiguë myéloblastique[2] (AML). C'est la première étude publiée qui propose une méthode de prédiction basée sur les biopuces.

[1]Leucémie aiguë caractérisée par la prolifération incontrôlée de lymphocytes immatures dans le sang et la moelle

[2]La leucémie aiguë myéloblastique est causée par un surnombre d'un autre type de cellules, les ńgranulocytesż. Ces cellules demeurent immatures et affectent les globules blancs, les globules rouges et les plaquettes de la même manière que les lymphocytes dans les cas de leucémies aiguës lymphoblastiques

En 2000, Alizadeh *et al.* [7] mettent en évidence l'existence de deux catégories de Démembrement des lymphomes à grandes cellules B (DLBCL)[1] avec des signatures distinctes, l'une qui correspondrait au profil d'expression des cellules B normales des centres germinatifs, et l'autre à celui des cellules B du sang périphériques. La survie à 5 ans dans la première catégorie est meilleure que dans la seconde, cette stratification n'étant pas possible avec l'IPI[2] seul. Cette étude a été d'une grande importance en hématologie. Plus tard, Shipp *et al.* [8] identifient un ensemble de 13 gènes jugés suffisant pour prédire la survie à cinq ans de patients atteints de DLBCL.

Une autre étude fréquemment citée a été effectuée par Van't Veer *et al.* [9], puis reprise par Vijver *et al.* [10] pour le cancer du sein. Cette équipe a identifié un profil génétique constitué de 70 gènes permettant de prédire la survenue ou non de métastases à cinq ans chez des patientes atteintes d'un cancer du sein en présence de ganglions axillaires. Selon les auteurs, ces 70 gènes permettent d'améliorer la prédiction donnée par les critères cliniques classiques, tels que le grade histologique ou le stade d'envahissement ganglionnaire. Si dans un premier temps, les études ont eu pour objectif de relier l'expression des gènes à un phénotype en classes (types ou sous-types de cancer, groupes de bons ou mauvais pronostic, etc), les cliniciens se sont ensuite intéressés à relier l'expression des gènes à des données de survie, pour étudier les délais de survie globale ou de survenue d'une rechute.

[1] Diffuse Large B Cell Lymphoma en anglais. Le DLBCL est un des sous-types les plus fréquents des lymphomes malins, représentant 30 à 40 % des lymphomes de l'adulte.

[2] L'IPI est un outil clinique développé par les oncologues pour aider à prédire le pronostic de patients atteints d'un lymphome non Hodgkinien agressif. Cet index intègre les critères suivants : l'âge (>60 ans ou non), le statut de la maladie, le nombre de sites extra-nodaux, le taux de LDH, et le niveau d'état de santé général.

Chapitre 2

Contexte méthodologique

Toute cette partie ne se veut pas exhaustive. La littérature sur les méthodes d'analyse des biopuces étant très vaste, seules celles principalement évoquées seront citées ici.

2.1 Identification de gènes marqueurs : méthodes d'analyse différentielle

2.1.1 Tests statistiques

Le but d'un test statistique est de tester une hypothèse définie par une question, ici biologique, définie à priori. On définit pour cela une hypothèse nulle H_0, et une hypothèse alternative, H_1. Les tests statistiques conduisent à deux types d'erreurs :1- Rejet à tort de l'hypothèse nulle, appelée erreur de première espèce ou risque de type I, et notée α ; 2- Acceptation à tort de l'hypothèse nulle , appelée erreur de seconde espèce ou risque de type II, et notée β. L'erreur de seconde espèce est reliée à la puissance du test $(1-\beta)$, qui est la probabilité de rejeter H_0 à raison. La p-value correspond à la probabilité d'obtenir, sous H_0, une valeur de la statistique de test T supérieure ou égale à celle observée t : $p(|T| \geq t|H_0)$. La statistique de test est transformée en une variable suivant une loi uniforme sur $[0,1]$ sous H_0.

Dans le cas des biopuces, on teste simultanément autant d'hypothèses qu'il y a de gènes : on évalue simultanément pour chacun des j gènes, l'hypothèse nulle H_j de non association entre les niveaux d'expression X_j et un phénotype d'intérêt. La prise de décision pour chacune des p hypothèses testées génère quatre cas de figures résumés dans le tableau 2.1. U est le nombre de

		Decision		
		Ho	H1	Total
Vérité	Ho	U	V	p_0
	H1	T	S	p_1
	Total	1-R	R	p

Tab. 2.1: *Notations*

vrais négatifs (VN), V le nombre de faux positifs (FP), T le nombre de faux négatifs (FN), et S le nombre de vrais positifs (VP).

2.1.2 Erreur de type I

Considérons un test statistique quelconque pour lequel la valeur pour chaque gène j est T_j. Pour chacun des gènes, cette valeur T_j est évaluée par rapport à un seuil c_α dont la valeur dépend du risque α autorisé. Pour un gène donné, on rejette l'hypothèse nulle H_j en faveur de l'hypothèse alternative H_1 si $|T_j| \geq c_\alpha$. La probabilité de déclarer un test significatif à tort est alors α. Par suite, la probabilité de déclarer au moins un test significatif parmi p tests indépendants est $1-(1-\alpha)^p$. On obtient un risque global qui augmente avec le nombre d'hypothèses testées, d'où la nécessité de contrôler ce risque global en diminuant en conséquence le risque individuel de chacun des tests. Cela revient à corriger les p-values de chacun des tests ; on parle de p-values ajustées. Ces p-values ajustées sont évaluées chacune au seuil α.

Deux grandes familles de méthodes d'ajustement des p-values existent, donnant deux définitions différentes du risque global : celles basées sur le FWER (Family Wise Error Rate), et celles basées sur le FDR (False Discovery Rate). Dudoit et al. [11] proposent une étude détaillée de ces méthodes de correction.

2.1.2.1 Family Wise Error Rate (FWER)

Le FWER est la probabilité de rejeter à tort au moins l'une des hypothèses nulles testées, soit $FWER = pr(V > 0)$. Contrôler le FWER, par exemple au seuil de 5%, permet d'être confiant à 95% de n'avoir aucun faux positif.
Les procédures de correction du FWER peuvent être classées en trois catégories principales : les procédures en une seule étape (Bonferroni, Sidak, Westfall et Young), les procédures séquentielles descendantes (Holm, Sidak SD, Westfall et Young SD) ou ascendantes (Hochberg). Les

méthodes de Westfall et Young présentent la particularité d'être basées sur une permutation des classes des individus, permettant ainsi de tenir compte de la structure de corrélation qui peut exister entre les gènes.

Globalement, ce sont des procédures très conservatrices, c'est à dire que peu de gènes sont sélectionnés. Elles sont plutôt utilisées dans le cadre d'analyses décisionnelles, quand le coût des erreurs de type I est élevé.

2.1.2.2 False Discovery Rate (FDR)

Le FDR, proposé pour la première fois par Benjamini et Hochberg [12], correspond à la proportion attendue de faux positifs parmi les gènes déclarés significatifs, soit en reprenant les notations du tableau 2.1, $FDR = E(V/R)$. Contrôler le FDR au seuil de 5% permet d'affirmer qu'en moyenne, le taux de faux positifs est inférieur à 5%.
Les procédures de correction du FDR peuvent également être classées en trois grandes catégories : les procédures séquentielles ascendantes (Benjamini et Hochberg, Benjamini et Yekutieli), les procédures "en deux étapes" (Benjamini et al.) et les procédures "adaptatives" (Benjamini et Hochberg).

Reiner *et al.* proposent une revue de ces différentes procédures [13]. Elles sont privilégiées dans le cadre d'analyses exploratoires, ce qui est le plus souvent le cas des analyses de puces à ADN.

Storey *et al.* [14] proposent la notion de q-value, qui mesure pour chaque gène la proportion moyenne de faux positifs estimée parmi tous les gènes plus significatifs que le gène considéré.

Le FDR local, plus précis, donne pour chaque gène sa probabilité d'être un faux positif. Différentes approches ont été proposées pour le calculer [15, 16].

2.1.3 Erreur de type II

Le contrôle du risque de première espèce n'est pas suffisant ; il est également nécessaire d'introduire le risque de seconde espèce à travers la puissance. Avec les notations du tableau 2.1, le risque de seconde espèce est défini par $\beta = E(T)/p1$. Par suite, la puissance est définie par $(1 - \beta) = E(V)/p1$. La puissance et le nombre de sujets nécessaires pour détecter une différence pré-définie sont étroitement liés : la puissance d'un test est d'autant plus grande que la taille de l'échantillon est élevée. Pour les premières analyses de biopuces, la question de

la puissance a d'abord été occultée, mais elle a fait l'objet de travaux récents qui proposent différentes approches pour le calcul du nombre de sujets nécessaires dans le cas de tests multiples ou de méthodes de prédiction [17, 18, 19, 20, 21, 22, 23, 20, 24].

2.2 Identification de nouvelles classes de tumeurs : méthodes d'analyse non supervisée

Méthode de clusters hiérarchiques. L'objectif de cette méthode est de regrouper dans un même groupe, ou cluster, des entités proches, ici les gènes ou les individus. Ces groupes sont construits itérativement en regroupant à chaque itération les entités, puis les groupes d'entités, les plus proches. Un arbre de classification, ou dendrogramme, est ainsi construit, à la racine duquel toutes les entités sont finalement regroupées. Deux mesures de similarité doivent être définies pour construire cet arbre. Elles définissent deux types de distances :

- Une distance entre les entités, la distance euclidienne, ou la corrélation par exemple.
- Une distance entre les groupes constitués. Cette distance peut être mesurée par exemple par les méthodes de :
  - Saut minimum (ou single linkage en anglais) : la distance entre deux groupes est donnée par la distance minimum entre les couples d'entités de ces groupes.
  - Saut maximum (ou complete linkage en anglais) : la distance entre deux groupes est donnée par la distance maximum entre les couples d'entités de ces groupes.
  - Saut moyen (ou average linkage en anglais) : la distance entre deux groupes est donnée par la distance moyenne entre tous les couples d'entités de ces groupes.

Eisen *et al.* [25] se servent de cette méthode pour proposer un mode d'interprétation visuelle des données de biopuces. Ils proposent une méthode de classification hiérarchique couplée à une colorisation du tableau de données qui met en évidence les proximités entre gènes et entre individus respectivement, en permettant une double classification des gènes et des individus. La métrique qui définit ces proximités est basée sur la corrélation. La figure 2.2, issue des résultats d'une étude conduite sur le lymphome par Alizadeh *et al.* [**?**], illustre cette méthode.

Sur la gauche est représenté le dendrogramme pour les gènes, et sur le haut le dendrogramme des patients. Cette visualisation permet de regrouper les patients partageant des sous-types de pathologies communes. Parallèlement, les réseaux de gènes dont les niveaux d'expression sont

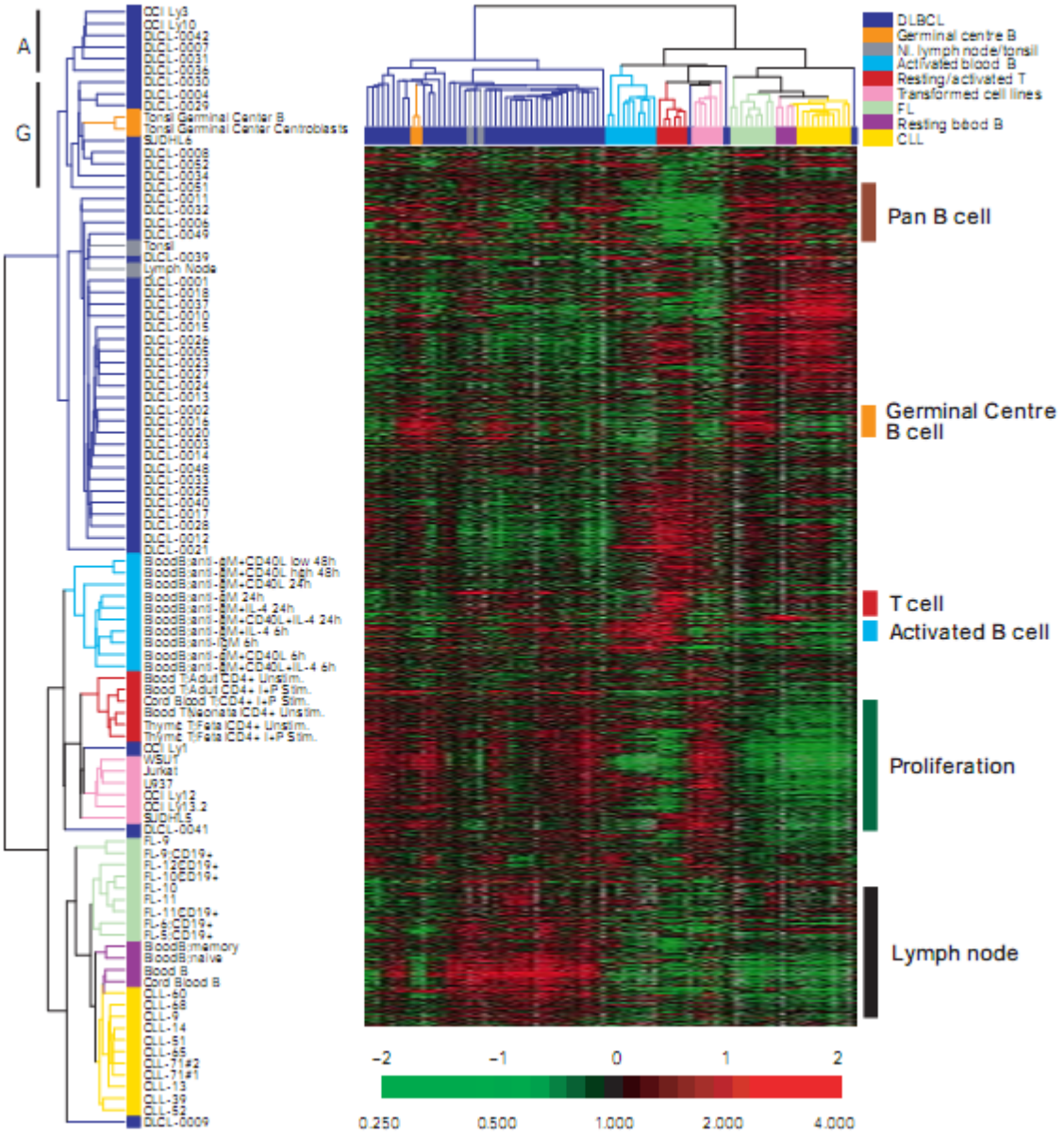


FIG. 2.1: *Exemple de visualisation de clusters hiérarchiques issu d'une étude sur le lymphome d'Alizadeh et al.*

propres à chacun de ces sous-types sont identifiés. Les zones rouges et vertes correspondent respectivement à une sur et sous-expresssion des gènes considérés.

Méthode de réallocation dynamique, dite des k-means. Cette méthode regroupe cette fois les entités selon un procédé non hiérarchique [26]. On en trouve un exemple d'application à l'analyse des biopuces par Herwig *et al.* [27]. Le principe de la méthode est de trouver une partition des individus ou des gènes dont le profil est similaire, en un nombre k préalablement défini de groupes. Les groupes sont tels que la distance entre leurs centroïdes[1] est minimisée, selon une métrique particulière choisie. Si l'objectif est par exemple d'identifier des groupes d'individus partageant le même profil d'expression génétique, l'algorithme est le suivant :

1. k points sont placés aléatoirement dans l'espace des gènes. Ce sont les centroïdes initiaux correspondant aux k groupes.
2. Chaque individu est ensuite attribué au groupe dont le centroïde est le plus proche selon une métrique prédéfinie.
3. Une fois l'ensemble des individus ainsi répartis dans les k groupes, les positions des k centroïdes sont recalculées.
4. Les étapes 2 et 3 sont réitérées jusqu'à ce que les positions des centroïdes restent fixes.

On peut noter que les résultats de cet algorithme sont sensibles à la position des k centroïdes initiaux.

Méthode d'Analyse en Composantes Principales (ACP). Elle est basée sur les méthodes d'analyse factorielle. En définissant le sous-espace de projection dans lequel la variabilité entre les entités est maximale, elle permet de visualiser les entités proches, gènes ou individus [28]. On en trouve un exemple d'application à l'analyse des biopuces par Alter *et al.* ou Landgrebe *et al.* [29, 30]. Cette méthode sera davantage approfondie dans la partie 4.

2.3 Classement d'une tumeur parmi des classes connues : méthodes d'analyse supervisée

L'objectif est de construire un modèle prédictif capable de prédire au mieux le phénotype de nouveaux individus à partir de leurs données d'expression.

[1] Le centroïde d'un groupe correspond à l'isobarycentre du nuage de points des individus dans l'espace des gènes.

Avant même de choisir un modèle, il peut être utile de savoir si les données sont suffisamment informatives. Dans cet objectif, Goeman *et al.* [31, 32] ont proposé un test global qui permet de tester l'associativité entre le niveau d'expression des gènes et la réponse clinique d'intérêt. Le principe est de tester globalement si les patients qui ont un profil d'expression similaire ont également une réponse similaire. La réponse peut être de type facteur ou données de survie.
Le test repose sur le calcul d'une statistique de test notée Q. Pour p gènes, l'hypothèse nulle du test global d'association entre les données d'expression et la réponse est définie par H_0 : $\beta_1 = ... = \beta_j = \beta_p$. Les paramètres $\{\beta_j\}_{j=1}^{p}$ sont les paramètres du modèle considéré, modèle de régression linéaire généralisé dans le cas d'une réponse de type facteur, et modèle de Cox dans le cas d'une réponse de type survie. En considérant que les β_j sont issus d'une même distribution de moyenne nulle et de variance τ^2, on peut réécrire l'hypothèse nulle comme $H_0 : \tau^2 = 0$. La statistique Q permet de traduire cette hypothèse. Si le test associé à cette statistique est significatif, la recherche d'un modèle prédictif est justifiée. Au contraire, un test non significatif laisse supposer qu'il y a peu de gènes différentiels, et qu'il y a peu d'espoirs de trouver un modèle prédictif intéressant.

2.3.1 Méthodes d'analyse proposées dans la littérature.

Méthodes basées sur des proximités locales. La méthode la plus simple, non paramétrique, est celle des k plus proches voisins (knn pour k Nearest Neighbors en anglais) [33]. Pour prédire le phénotype d'un nouvel individu, cette méthode consiste à prendre en compte les k individus du jeu d'apprentissage les plus proches de cet individu selon une métrique définie, et de lui attribuer le phénotype le plus représenté parmi ses k plus proches voisins. Le nombre de voisins optimal à prendre en compte est déterminé par validation croisée.

La méthode des centroïdes est très proche de cette méthode. Cette fois, chaque groupe est représenté par son centroïde et un nouvel individu est classé dans le groupe dont le centroïde est le plus proche. La méthode de "nearest shrunken centroid"[1] de Tibshirani *et al.* en est une variante [34]. Dans cette variante, les coordonnées du centroïde de chaque classe sont réduites vers zéro par soustraction d'une valeur seuil qui dépend du centroïde commun à toutes les classes. Cette opération de seuillage permet de limiter les effets dûs au bruit et de sélectionner des gènes : si la valeur d'un gène est mise à zéro dans toutes les classes, il ne contribue plus au

[1] On trouve aussi cette méthode sous le nom de PAM, pour Prediction Analysis for Microarray.

classement d'un nouvel individu.

Méthode SVM (Support Vector Machine). Cette méthode [35, 36] recherche des hyperplans[1] dans lesquels une séparation linéaire optimale permet de distinguer les groupes d'individus. Dans certains cas en effet, il n'existe pas de séparateur linéaire permettant de séparer les groupes d'individus. L'idée des SVM est alors de reconsidérer le problème dans un espace de dimension supérieure dans lequel il existera un séparateur linéaire permettant de classer les individus. Le séparateur est ensuite projeté dans l'espace d'origine pour visualiser les résultats des classements.

La construction de l'hyperplan optimal est basée sur un algorithme itératif par minimisation d'une fonction d'erreur définie au préalable.

Méthodes d'analyse discriminante. La méthode d'analyse discriminante recherche une combinaison linéaire des gènes qui rende simultanément maximale la distance entre les groupes et minimale les distances entre individus d'un même groupe.

Soient $\mathbf{x} = (x_1, ..., x_p)$ les niveaux d'expression des p gènes d'un individu i, et $\{\mu_k; k = 1, ..., m\}$ les profils génétiques moyens des m groupes. Soit Σ_k la matrice de variance-covariance du groupe k.

La règle de décision de l'analyse discriminante revient à affecter un individu dont le profil génétique est décrit par $\mathbf{x}$ à la classe dont le profil génétique moyen est le plus proche. En faisant l'hypothèse que les niveaux d'expression des gènes suivent une loi normale multivariée $(\mathbf{x}|y = k \sim N(\mu_k, \Sigma_k))$, ceci revient à minimiser la quantité $\left[(\mathbf{x} - \mu_k)\Sigma_k^{-1}(\mathbf{x} - \mu_k)' + log\,|\Sigma_k|\right]$.

Dans le cas où la matrice $\Sigma_k = diag(\sigma_{k1}^2, ..., \sigma_{kp}^2)$ est diagonale, on parle d'analyse discriminante diagonale quadratique (DQDA). Dans le cas où cette matrice diagonale est commune aux k classes, $\Sigma_k = \Sigma = diag(\sigma_1^2, ..., \sigma_p^2)$, on parle d'analyse discriminante diagonale linéaire (DLDA). Les vecteurs de moyenne ainsi que la matrice de variance-covariance sont estimés sur le jeu de données sur lequel le modèle est construit ; on notera $\hat{\mu}_k = \bar{\mathbf{x}}_k$.

[1] Dans un espace de dimension finie n, les hyperplans sont les sous-espaces vectoriel de dimension n-1

Dans le cas de l'analyse discriminante diagonale linéaire, et dans le cas particulier de deux classes, la règle de décision permet d'affecter un individu à la classe 1 si

$$\sum_{j=1}^{p} \frac{(\bar{x}_{1j} - \bar{x}_{2j})}{\hat{\sigma}_j^2} \left(x_j - \frac{(\bar{x}_{1j} - \bar{x}_{2j})}{2} \right) \geq 0$$

Dans la première analyse supervisée de biopuces, Golub *et al.* utilisent une méthode de vote pondéré[1], où chaque gène vote pour l'attribution d'un nouvel individu à l'une ou l'autre des classes considérées [6, 8, 37]. Dudoit *et al.* [38] montrent que cette méthode est une variante de l'analyse discriminante où $\left[(\bar{x}_{1j} - \bar{x}_{2j})/\hat{\sigma}_j^2\right]$ est remplacé par $[(\bar{x}_{1j} - \bar{x}_{2j})/(\hat{\sigma}_{1j} + \hat{\sigma}_{2j})]$.

Dans le même article, les auteurs proposent une revue approfondie des différents types d'analyses discriminantes.

La méthode d'analyse inter-groupes repose sur les mêmes principes que la méthode d'analyse discriminante. Elle a été utilisée dans le contexte des biopuces par Culhane *et al.* [39]. Baty *et al.* [40] proposent d'inclure dans la méthode une procédure de sélection de gènes basée sur le jackknife, qui permet d'améliorer les capacités prédictives de la méthode initiale. En conservant les gènes qui contribuent le plus à la construction des axes discriminants, c'est à dire ceux qui ont les plus forts coefficients dans la combinaison linéaire définissant les axes discriminants, l'analyse inter-groupes peut également être employée comme méthode de sélection de gènes.

Les analyses discriminantes et inter-groupes seront redéveloppées plus en détail dans la partie 4.

Méthode de forêts aléatoires. Plus récemment, la méthode de forêts aléatoires[2] a été appliquée dans le cadre de l'analyse des biopuces [41]. Cette technique a pour objectif de réduire la variabilité du prédicteur obtenu par les arbres binaires de classification (CART)[3] [42], très instables, en combinant leurs résultats. La méthode CART est basée sur un découpage, par des hyperplans, de l'espace engendré par les variables.

Une forêt aléatoire consiste en un nombre arbitraire d'arbres simples, utilisés pour calculer un vote pour la classe la plus représentée (classification), ou dont les réponses sont combinées pour obtenir une estimation de la variable dépendante (régression). Un sous-ensemble de pré-

[1] Weighted Voting en anglais
[2] Random Forest en anglais
[3] Classification And Regression Trees en anglais

dicteurs est choisi indépendamment pour chacun des arbres de la forêt. Les forêts aléatoires sont une des techniques d'agrégation de modèles, dont le but est de diminuer l'erreur de prédiction. Elles sont une adaptation aux arbres de classification binaires de la méthode de bagging[1] [43]. Cette méthode adopte une stratégie aléatoire en moyennant les prédictions obtenues sur un nombre déterminé d'échantillons bootstrap. Une alternative au bagging est le boosting, un algorithme adaptatif où chaque nouvel arbre est une version adaptative du précédent en donnant plus de poids, lors de l'estimation suivante, aux observations mal ajustées [44].

Pour une approche comparative des méthodes citées, on pourra se reporter entre autres aux travaux de Romualdi *et al.* [45] (LDA,SVM, knn,PAM), Dudoit *et al.* [38] (LDA, QDA, knn, méthodes d'agrégation de modèles), ou encore Boulesteix (SVM, PAM, knn, Analyse discriminante précédée de PLS) [46]. La conclusion générale qui résulte de ces comparaisons est qu'il n'existe pas une méthode réellement meilleure que les autres en toute situation. Ces méthodes donnent des résultats globalement équivalents, et l'une peut prendre le dessus sur l'autre pour un jeu de données particulier. Selon Dudoit *et al.* [38], les méthodes les plus simples, telles que la méthode des k plus proches voisins, sont parfois même plus performantes que des méthodes complexes, car elles font moins d'hypothèses sur la structure des données. Un modèle plus simple est par ailleurs moins sujet au sur-ajustement.

Adaptation aux données de survie. Pour l'analyse des données de survie, les premiers travaux utilisaient le principe suivant : 1- découverte de classes de patients par méthodes de classification non supervisée ; 2- construction de courbes de Kaplan-Meyer dans chacune des classes identifiées, avec mise en évidence d'une différence de survie entre les groupes [47]. Une autre approche était de construire des clusters de gènes et d'introduire les profils moyens correspondant à ces clusters dans un modèle de Cox. L'inconvénient majeur de ces procédés est qu'ils ne tiennent pas compte des informations sur la survie pour la génération des classes de gènes ou de patients. Plus tard, des adaptations du modèle de Cox ont été proposées, comme décrit dans la partie 2.3.3.

Particularité des données de biopuces. Ce qui caractérise les données de biopuces est que le nombre d'individus n est très supérieur au nombre de gènes p. Cette particularité rend l'utilisation des méthodes citées ci-dessus difficile, voire impossible quand elles nécessitent l'esti-

[1] Bootstrap Aggregating

mation de la matrice de variance-covariance de la matrice de données X comme c'est le cas par exemple pour l'analyse discriminante. On parle de "fléau de la dimension". Le même problème intervient dans le modèle de Cox. Quel que soit le modèle de régression, sa construction dans le cas où le nombre de variables est supérieur au nombre d'individus pose un double problème, celui de la multi-colinéarité entre les variables et celui du sur-ajustement. Ce phénomène apparaît quand un modèle prédictif complexe (i.e avec beaucoup de paramètres) est construit sur un jeu de données trop petit. Le modèle s'ajuste trop aux données, y compris au bruit, ce qui limite les performances de la prédiction sur de nouvelles données. L'évaluation de la qualité prédictive du modèle, capacité à prédire sur de nouveaux jeux de données, sur le jeu de données sur lequel il a été construit, est surestimée ; le modèle est optimiste.

Trois solutions permettent de résoudre ce double problème, décrites ci-dessous : la sélection ou l'extraction de variables, regroupées sous le terme de réduction de la dimension, et la régression parcimonieuse. Si les deux premières solutions permettent essentiellement de résoudre le problème de la muti-colinéarité, la dernière solution a plus particulièrement pour objectif de réduire l'optimisme des modèles générés.

2.3.2 Réduction de la dimension

Pour contourner le problème de multi-colinéarité, il est nécessaire de travailler dans un espace des gènes de dimension inférieure, en utilisant au préalable une méthode dite de réduction de la dimension. D'un point de vue terminologique, le terme de réduction de la dimension désigne le fait de réduire le nombre de variables à considérer. Il comprend deux types de démarche : la sélection de variables et l'extraction de variables. La première démarche vise à sélectionner un sous-ensemble de variables, tandis que la seconde vise à projeter les entités dans un espace de dimension inférieure où les nouvelles variables, les composantes, sont des combinaisons linéaires des anciennes. Dans le domaine du transcriptome, le terme "réduction de la dimension" est associé à cette dernière démarche.

2.3.2.1 Sélection univariée de variables

L'objectif est de sélectionner un nombre restreint de gènes en se limitant à ceux dont l'effet est le plus marqué. Ces gène sont dits différentiels. Pour cela, un score est calculé pour chacun des gènes, et les gènes avec les meilleurs scores sont retenus. Parmi ces scores, et dans le cas où les individus sont répartis en deux groupes, la statistique du t de Student est la plus souvent

utilisée, ou encore la statistique de Welch qui permet de tenir compte du fait que les variances des niveaux d'expression des gènes ne sont pas les mêmes dans les deux groupes.
Pour être en mesure de comparer les scores obtenus pour chacun des gènes, la distribution de ces scores doit être indépendante du niveau d'expression des gènes. Or, pour des gènes peu exprimés, la variance peut être faible, ce qui conduit à un score artificiellement élevé. Pour tenir compte de ce phénomène, Tusher *et al* [48] proposent d'ajouter une constante positive au dénominateur de la statistique de t. Pour choisir la valeur de cette constante, les statistiques de test sont réparties en groupes de variance homogène. La constante est choisie de telle sorte que la dispersion de la valeur du test statistique ne varie pas d'un groupe à l'autre.
Un autre type de méthode, non paramétrique, de sélection de gènes est basée sur la méthode de "Produit des rangs", ou "Rank Product" en anglais [49]. Pour chacun des individus, les gènes sont ordonnés sur la base d'un score prédéfini, l'inverse du coefficient de variation par exemple. Un gène placé en première position pour les n sujets est différentiellement exprimé pourra alors être supposé différentiel. Partant de ce principe, le produit des rangs occupés par chaque gène g pour chacun des n individus est calculé comme suit :

$$RP_g = \prod_{i=1}^{n} (r_{i,g}/p)$$

où $r_{i,g}$ est le rang du gène g chez l'individu i. Les gènes sélectionnés sont ceux pour lesquels les valeurs de RP sont les plus faibles.

Jeffery *et al* [50] proposent une comparaison détaillée des principales méthodes de sélection employées pour l'analyse des biopuces. Ils montrent que le choix de ces méthodes dépend de la variance intra-groupes estimée sur le jeu de données à analyser.

Dans le cas de données censurées, l'étape de sélection univariée consiste généralement à sélectionner les gènes pour lesquels les tests du log-rank sont les plus significatifs.

Le but étant uniquement de sélectionner un sous-ensemble de gènes, il n'y a pas lieu de tenir compte du degré de significativité des gènes, ni d'apporter de correction due à la multiplicité des tests.

Du fait de la sélection univariée, les corrélations qui existent entre les gènes ne sont pas prises en compte. Or, ces corrélations peuvent introduire des effets combinés sur le phénotype

considéré, d'où l'intérêt d'en tenir compte, comme c'est le cas par exemple dans les méthodes d'extraction de variables.

2.3.2.2 Extraction de variables

Les deux méthodes les plus fréquemment utilisées sont l'Analyse en Composantes Principales (ACP) et la méthode Partial Least Squares (PLS).

Méthode d'Analyse en Composantes Principales (ACP). L'ACP recherche l'espace de projection qui maximise la variabilité totale entre les individus. Les composantes $\mathbf{w}_j$, $\{j = 1, .., k\}$, axes du nouvel espace ainsi obtenu, sont telles que : $\mathbf{w}_k = argmax_{\mathbf{w'w=1}} var(X\mathbf{w}, \mathbf{y})$, avec la contrainte : $\mathbf{w}'\Sigma\mathbf{w}_j = 0$ pour tout $1 \leq j < k$, où $\Sigma = X'X$. Cette méthode sera approfondie dans la partie 4.

Méthode de Partial Least Squares. Initialement développée dans le domaine de la chimiométrie [51], la méthode Partial Least Square (PLS) est une approche de réduction de la dimension couplée à un modèle de régression. Elle recherche le sous-espace de projection qui maximise la covariance entre les niveaux d'expression et la covariable d'intérêt, définie par le vecteur $\mathbf{y}$. Les composantes $\mathbf{w}_j$, $\{j = 1, .., k\}$, qui définissent ce sous-espace sont telles que $\mathbf{w}_k = argmax_{\mathbf{w'w=1}} cov(X\mathbf{w}, \mathbf{y})$ avec la contrainte : $\mathbf{w}'\Sigma\mathbf{w}_j = 0$ pour tout $1 \leq j < k$.

La contrainte peut répondre à des critères différents, qui conduisent à des algorithmes différents. La contrainte exposée ci-dessus correspond aux algorithmes PLS1 et SIMPLS qui prennent en compte une réponse respectivement univariée et multivariée. La méthode PLS permet également de sélectionner des gènes d'intérêt à partir des coefficients des gènes sur les composantes. Boulesteix a montré que l'ordre des coefficients des gènes sur la première composante correspond à celui de tests de F [46]. Elle a également montré que dans le cas de deux groupes, la première composante obtenue par l'algorithme SIMPLS de la PLS coïncide avec l'axe discriminant obtenu par l'analyse inter-groupes [46].

La méthode PLS a été adaptée aux données de survie dans le cadre des biopuces de différentes manières. Nguyen et Rocke proposent une méthode en deux étapes : dans un premier temps les composantes PLS sont extraites en utilisant comme réponse les délais de survie sans tenir compte de la censure [52, 53], puis elles sont introduites comme variables dans un modèle de Cox. Plus tard, cette approche est améliorée par d'autres auteurs pour tenir compte de la nature particulière des données de survie dans la réduction de la dimension [54, 55, 56].

Boulesteix et Strimmer [57] ont reconsidéré la théorie sous-jacente de la PLS et proposent une revue complète de son application à l'analyse du transcriptome.

Autres approches Il existe d'autres méthodes de réduction de la dimension moins fréquemment rencontrées dans la littérature, issues de la théorie de la SDR (Sufficient Dimension Reduction) : SIR (Sliced Inverse Regression) [58] et MAVE (Minimum Average Variance Estimation), qui étend la première approche [59]. Antoniadis *et al.* [60] et Chiaromonte *et al.* [61] montrent respectivement l'intérêt de la première et seconde méthode dans le contexte des biopuces. Une particularité des méthodes MAVE et SIR est qu'elles nécessitent de sélectionner dans un premier temps un sous-ensemble de gènes. Li et Li [62, 63] ont adapté la méthode SIR aux données censurées.

Bair et Tishirani [64] proposent une méthode dite "semi-supervisée" dans le but d'identifier des sous-groupes biologiques (type de tumeurs par exemple), qui aient également un sens en terme de survie. Pour cela ils développent une analyse en composantes principales supervisée qui consiste à effectuer une ACP en se basant uniquement sur le sous-ensemble de gènes les plus corrélés à la survie (scores de risque des modèles de Cox univariés les plus significatifs). Ce sont les composantes qui sont ensuite intégrés dans le modèle de Cox final.

2.3.3 Méthodes de régression parcimonieuse

Dans le cadre des biopuces, ces méthodes ont surtout été utilisées pour adapter le modèle de Cox au "fléau de la dimension". Ces méthodes dites de régression parcimonieuse, ou de régularisation, sont des méthodes de maximisation de la vraisemblance sous contrainte : elles reposent sur une pénalisation de la vraisemblance du modèle, qui a pour conséquence de réduire vers zéro les coefficients des variables qui ne contribuent pas ou peu à la prédiction. On parle de "rétrécissement"[1] des coefficients. Le paramètre de pénalisation, λ, est celui qui optimise un critère d'évaluation de la qualité des modèles déterminé au préalable. Différents types de pénalité ont été proposés dans la littérature.

Pénalité de type $L2$. La pénalité de type $L2$ contraint les paramètres à être tels que $\left\{\left(\sum_{j=1}^{p} \|\beta_j\|^2 \leq \lambda\right)\right\}$, où λ est le paramètre de pénalité. On parle également de "régression ridge" pour désigner ce type de pénalisation.

[1] Traduction littérale de "shrinkage" en anglais.

Eilers *et al.* [65] ont appliqué la régression ridge au modèle logistique et démontré sa performance sur le jeu de données de Golub déjà évoqué (patients ALL vs AML) [6]. Le paramètre de pénalité λ qui maximise l'Akaike's Information Criterion (AIC) est obtenu par validation croisée. L'AIC est une mesure de la qualité d'ajustement d'un modèle permettant un compromis entre la complexité du modèle et son ajustement aux données. Il est défini par $(-2l(\beta)+2k)$, où $l(\beta)$ est la log-vraisemblance du modèle, et k le nombre de paramètres retenus dans ce modèle.

Houwelingen *et al.* [66] ont adapté cette méthode aux données de survie. Cette fois, le paramètre λ retenu est celui qui maximise la vraisemblance partielle cross-validée introduite par Verweij et Houwelingen [67]. Celle-ci est définie par $\sum_{i=1}^{n}\left[pl(\hat{\beta}^{(-i)}) - pl^{(-i)}(\hat{\beta}^{(-i)})\right]$, où $pl^{(-i)}$ est la log-vraisemblance partielle du modèle obtenu sur l'ensemble des $(n-1)$ patients privés du patient i, les paramètres $\hat{\beta}^{(-i)}$ maximisant $pl^{(-i)}(\beta)$.

Pawitan *et al.* [68] utilisent la réécriture du modèle de Cox comme un modèle de régression de Poisson, et introduisent une pénalisation de type $L2$ dans ce dernier modèle. ce modèle leur permet également d'introduire un effet aléatoire sur le niveau d'expression des gènes.

Un inconvénient de cette méthode de pénalisation est qu'elle ne permet pas de sélectionner directement des gènes d'intérêt : toutes les variables ont un coefficient non nul dans le modèle final.

Pénalité de type $L1$. La pénalité de type $L1$ contraint les paramètres à être tels que $\left\{(\sum_{j=1}^{p}|\beta_j| \leq \lambda)\right\}$. La méthode du Lasso, développée puis adaptée à la survie par Tibshirani, utilise cette norme [69, 70]. Elle n'est utilisable qu'à la condition que le nombre de variables reste inférieur au nombre de patients.

Le LARS (Least Angle Regression) [71] généralise le Lasso et permet de s'adapter au cas où $n << p$. Gui et Li [72] étendent le LARS au cas des données censurées. Les temps de calcul de leur approche sont cependant longs, ce qui a incité Segal [73] à proposer une modification de leur algorithme.

Dans le cas de variables très corrélées, seule une variable du cluster de gènes correspondant est sélectionnée. Par ailleurs, l'utilisation du LARS ne permet pas de sélectionner plus de variables qu'il n'y a de patients inclus dans l'analyse.

Autres approches dérivées. Ceci est rendu possible dans la méthode du gradient proposée par Friedman [74]. Cette fois, les coefficients sont estimés de manière itérative en se déplaçant

dans la direction opposée au gradient de la log-vraisemblance. Cette méthode permet d'approximer les résultats obtenus par les deux types de pénalisation décrits ci-dessus. Gui et Li ont adapté cette méthode au modèle de Cox [75] ; ceci sera décrit plus en détail dans la partie 5.

Li et Luan [76] proposent une généralisation de la méthode SVM pour les données de survie [77]. Ils se basent sur le fait que la méthode SVM peut être réécrite comme une méthode de pénalisation en se plaçant dans un espace de Hilbert. Dans ce cas on cherche à optimiser directement la fonction de risque plutôt que d'optimiser la valeur des paramètres β.

Dans le même contexte théorique, les auteurs proposent une approximation de la fonction de risque basée sur une optimisation dans l'espace des fonctions plutôt que dans celui des paramètres [76]. Ils utilisent pour cela une méthode d'estimation non paramétrique de la fonction de risque basée sur la méthode de "gradient boosting machine" proposée par Friedman [78]. Cette méthode, proche de la méthode du gradient évoquée ci-dessus, aboutit à l'obtention d'une forme fonctionnelle propre à chaque gène. Selon les auteurs, un avantage de cette méthode est qu'elle s'affranchit de la contrainte forte de linéarité de la fonction de risque du modèle de Cox.

2.3.4 Validation des modèles

À terme, l'objectif pour le clinicien est de disposer de kits diagnostiques ou pronostiques. Mais avant l'utilisation en routine de ces tests en clinique, il est indispensable de valider les marqueurs détectés.

Une fois construit, le modèle prédictif doit en effet être évalué puis validé. Idéalement, la construction du modèle devrait être faite sur un jeu de données indépendant de celui sur lequel il sera évalué. En pratique, peu de jeux de données sont disponibles pour une même question biologique ; de ce fait, la construction et la validation du modèle se font sur un même jeu de données, divisé aléatoirement entre un jeu dit d'apprentissage ou de travail, et un jeu test. Le modèle est construit sur le premier, et l'erreur de prédiction est évaluée sur le second.

La validation croisée consiste à répéter ce processus un nombre déterminé s de fois ; l'erreur de prédiction finale est celle obtenue en moyenne sur les s étapes.
Il n'y a pas de règle fixe qui détermine les proportions relatives d'individus dans les jeux de travail et de test. Une solution est d'utiliser un processus dit de "Leave-One-Out Cross-Validation" (LOOCV), qui consiste à construire le modèle sur $(n-1)$ individus et à l'évaluer sur l'individu retiré. Il semble cependant plus adapté d'utiliser un processus "Leave-k-Out Cross-

Validation", qui consiste cette fois à enlever un pourcentage k de patients sur lequel évaluer le modèle.

Le modèle qui sera utilisé pour la prédiction sur un nouveau jeu de données est celui qui est construit à partir des n patients disponibles. L'objectif de la validation croisée est d'estimer l'erreur de prédiction qui serait faite sur un nouveau jeu de données.

Ambroise et McLachlan [79] puis Simon [80, 81] ont montré qu'il était indispensable d'introduire toutes les étapes de construction du modèle dans le processus de validation croisée, y compris s'il y a lieu l'étape de sélection de gènes. Les premiers illustrent ces propos par des études où cela n'avait pas été fait, et mettent en évidence le biais de sélection qui en découle. Le second met en garde contre des analyses ou des méthodes nouvelles publiées sans véritable validité statistique.

La notion de validité du modèle est étroitement liée avec celle de la validité des biomarqueurs détectés. Cette dernière peut s'exprimer en terme de reproductibilité : est-ce que le biomarqueur considéré est retrouvé dans des études différentes ? En reprenant l'analyse de sept études dont le but est de prédire l'issue d'un cancer, Michiels *et al.* [82] montrent que le sous-ensemble des gènes d'intérêt sélectionnés est très dépendant du jeu de données. Ils mettent en avant la nécessité de valider les résultats par validation croisée pour éviter des résultats trop optimistes. Ils montrent également qu'en utilisant des jeux de données d'effectifs plus importants, les résultats se stabilisent.

En se limitant au jeu de données sur le cancer du sein de Van't Veer [9], Ein-Dor *et al.* [83] aboutissent aux mêmes conclusions quant au manque de stabilité des signatures de gènes. Un peu plus tard, ils montrent que pour avoir des signatures reproductibles, la taille des études doit être augmentée [84].

Dans la continuité, Fan *et al.* [85] ont constaté que, toujours dans le cas du cancer du sein, différentes études conduites sur des jeux de données différents mènent à l'identification de profils pronostiques différents. En comparant les prédictions dérivées de cinq modèles différents sur un même jeu de données, ils ont observé que ces modèles sont particulièrement concordants en termes de prédiction et ont conclu que ce n'est pas la concordance des signatures, mais plutôt la concordance des prédictions qui doit être utilisée pour mesurer la reproductibilité.

Troisième partie

Développement méthodologique mis en oeuvre

Chapitre 3

Jeux de données

Les méthodes mises en oeuvre dans mon travail ont été appliquées sur des jeux de données réels ou simulés. Les jeux de données réels sont des jeux de données publics classiquement utilisés dans la littérature. Quant aux simulations, deux outils ont été développés, répondant chacun à un objectif différent. Le premier outil permet de simuler des données d'expression issues de deux groupes de patients, un groupe de patients malades et un groupe de patients sains par exemple. La particularité de ces simulations est qu'elles permettent de contrôler la structure des données. Le second outil permet de simuler deux types de variables : des variables clinico-biologiques classiques et des niveaux d'expression de gènes. Cette fois, le phénotype d'intérêt n'est pas un critère binaire mais des données de survie.

Nous présenterons dans un premier temps les jeux de données publics, puis dans un second temps les outils de simulation.

3.1 Jeux de données publics

Pour ces jeux de données, la réponse d'intérêt correspond à un phénotype binaire décrivant l'appartenance de chaque individu à l'une des deux classes. Le tableau 3.1 présente les jeux de données publics qui ont été utilisés.

Les jeux de données n^o 1 et 2 sont issus d'une étude de Shipp [8] concernant des patients atteints de lymphome à grandes cellules B (58 patients) ou de lymphomes folliculaires (19 patients). Les données d'expression proviennent d'une puce Affymetrix Hu6800 qui contient 7129 "probeset". Le jeu DLBCL.1 concerne uniquement les patients atteints du premier type de lymphome, répartis en deux groupes selon qu'il y a eu rechute/décès (26 patients), ou guérison

Numéro	Nom	Description des classes	Nb patients	Nb gènes
1	DLBCL.1	Guérison ou Rechute	58	6149
2	DLBCL.2	Folliculaire ou à Grandes Cellules B	77	7129
3	Prostate	Porteur ou non d'une tumeur	102	12625
4	Colon	Tumoral ou Normal	62	2000
5	Leucémie	ALL ou AML	72	7129
6	Myélome	Présence ou non d'une lésion lytique	173	12625
7	ALL.1	Origine B-cellulaire ou T-cellulaire	128	12625
8	ALL.2	Avec ou sans MDR	125	12625
9	ALL.3	Rechute ou non-rechute	100	12625
10	ALL.4	Avec ou sans translocation	95	12625

Tab. 3.1: *Jeux de données publics*

(32 patients) à cinq ans. Le jeu DLBCL.2 s'intéresse à la distinction entre les deux types de lymphomes. Les données sont disponibles sur le site internet du Broad Institute [86]. Elles ont été pré-traitées par la méthode RMA (Robust Multichip Average) [87], disponible dans le package *affy* de Bioconductor [88]. Ce pré-traitement consiste en trois étapes :1- un ajustement du bruit de fond, qui corrige l'intensité des perfect match (PM) puce par puce ;2- une normalisation par la méthode des quantiles [89] qui vise à imposer une distribution empirique des intensités commune à toutes les puces ;3- une mise en commun des intensités d'un même probeset, basée sur un algorithme itératif ("median polish algorithm" [90]) qui permet de combiner l'information provenant de l'ensemble des biopuces issues d'une même étude.

Le jeu de données n^o 3 est issu d'une étude de Singh *et al.* [91]. Il regroupe les niveaux d'expression de 12625 gènes (biopuces Affymetrix Hum95Av2) pour 102 patients répartis en deux groupes selon qu'ils sont ou non porteurs d'une tumeur de la prostate (respectivement 52 patients et 50 patients). Les données ont été téléchargées sur le site du Broad Institute [92]. Le même processus de prétraitement que pour les jeux de données concernant le DLBCL a été mis en oeuvre.

Le jeu de données n^o 4 provient d'une étude de Alon *et al.* [93]. Il est disponible dans la librairie *colonCA* de Bioconductor. Il donne les niveaux d'expression de 62 échantillons issus de patients qui souffrent d'un cancer du colon, dont 40 échantillons tumoraux et 22 échantillons normaux. Les puces utilisées pour l'analyse sont de type Affymetrix Hum6000. Les données ont été normalisées par la méthode des quantiles [89].

Le jeu de données n^o 5 provient d'une étude de Golub *et al.* [6]. Il est disponible dans la librairie *golubEsets* [94] de Bioconductor. Il fournit les niveaux d'expression de 7129 gènes (biopuces Affymetrix Hu6800) pour 72 patients, dont 47 souffrent de leucémie aiguë lymphoblas-

tique[1], et 25 de leucémie aiguë myéloblastique [2]. Les niveaux d'expression sont seuillés par 100 et 16000, qui correspondent aux seuils de détection et de saturation de l'appareil d'acquisition. Les données ont subis une transformation logarithmique en base 2.

Le jeu de données n^o 6 [95] fournit les niveaux d'expression de 12625 gènes (biopuce Affymetrix U95Av2) pour 173 patients atteints d'un myélome[3], parmi lesquels 36 présentent des lésions lytiques de l'os, et 137 non. Les données ont été téléchargées sur le site de Gene Expression Omnibus (GEO) [96] (numéro d'accession GDS531), puis pré-traitées par la méthode RMA.

Les jeux de données n^o 7 à 10 sont issus d'un même jeu de données analysé par Chiaretti *et al.* [97]. Les niveaux d'expression de 12625 gènes (biopuces Affymetrix hgu95av2) ont été analysés sur 128 patients atteints de leucémie aiguë lymphoblastique. Différentes covariables sont disponibles et les jeux n^o 7 à 10 correspondent à des sous-groupes de patients obtenus en considérant l'une ou l'autre de ces covariables.

- Le jeu n^o 7 sépare les patients selon que leur leucémie est à cellules B (95 patients) ou T (33 patients).
- Le jeu n^o 8 sépare les patients selon que leurs cellules présentent ou non une multirésistance aux médicaments (respectivement 24 et 101 patients).
- Le jeu n^o 9 sépare les patients selon qu'ils ont rechuté ou non dans un délai de deux ans (respectivement 65 patients et 35 patients).
- Le jeu n^o 10 sépare les patients selon qu'ils ont ou non la translocation $t(9;22)$[4] (respectivement 26 patients et 69 patients).

Les données sont disponibles après normalisation par la méthode gcRMA [98] dans la librairie *GOstats* de Bioconductor. Cette méthode de normalisation est une variante de la méthode RMA précédemment citée, qui tient compte lors de la correction du bruit de fond de la séquence nucléique des sondes.

[1] Leucémie aiguë caractérisée par la prolifération incontrôlée de lymphocytes immatures dans le sang et la moëlle

[2] La leucémie aiguë myéloblastique est causée par un surnombre d'un autre type de cellules, les "granulocytes". Ces cellules demeurent immatures et affectent les globules blancs, les globules rouges et les plaquettes de la même manière que les lymphocytes dans les cas de leucémies aiguës lymphoblastiques

[3] Le myélome est un cancer hématologique de la moelle osseuse. C'est une maladie caractérisée par le développement dans toutes les parties du squelette de multiples tumeurs ostéolytiques.

[4] La translocation t(9;22) correspond à un échange de segments entre les gènes ABL du chromosome 9 et BCR du chromosome 22. Cette translocation caractérise un sous-groupe des leucémies aiguës lymphoblastiques : les leucémies myéloïdes chroniques.

3.2 Jeux de données simulés

3.2.1 Premier outil de simulation

Le mode de simulation décrit ici permet de générer des jeux de données de structures variées, dans l'objectif de comprendre comment cette structure est prise en compte par les méthodes d'analyse discriminante précédée d'une ACP ou PLS, et d'analyse inter-groupes.

Puisque c'est la structure des données que nous voulions contrôler, nous avons commencé par la visualiser sur des jeux de données publics, en utilisant pour cela le premier plan d'une ACP intra-groupes. L'ACP intra-groupes a pour objectif de rechercher le sous-espace de projection dans lequel la variance intra-groupes est maximale. Ceci revient à rechercher le sous-espace de projection dans lequel la variabilité résiduelle après élimination de la variabilité due aux groupes est maximale. Dans notre cas, ceci revient à rechercher le sous-espace de projection qui maxime la variabilité résiduelle après prise en compte des différences de niveaux d'expression entre les groupes [99]. La structure de variance-covariance dans chacun des groupes peut ainsi être visualisée.

La figure 3.1 montre les résultats obtenus pour les jeux de données de Golub (cf tableau 3.1, jeu n^o 5) et de Shipp (jeu n^o 1). Les individus sont projetés dans le plan défini par les deux premières composantes de l'ACP intra-groupes.
Les deux jeux de données présentent des structures de configurations totalement différentes.

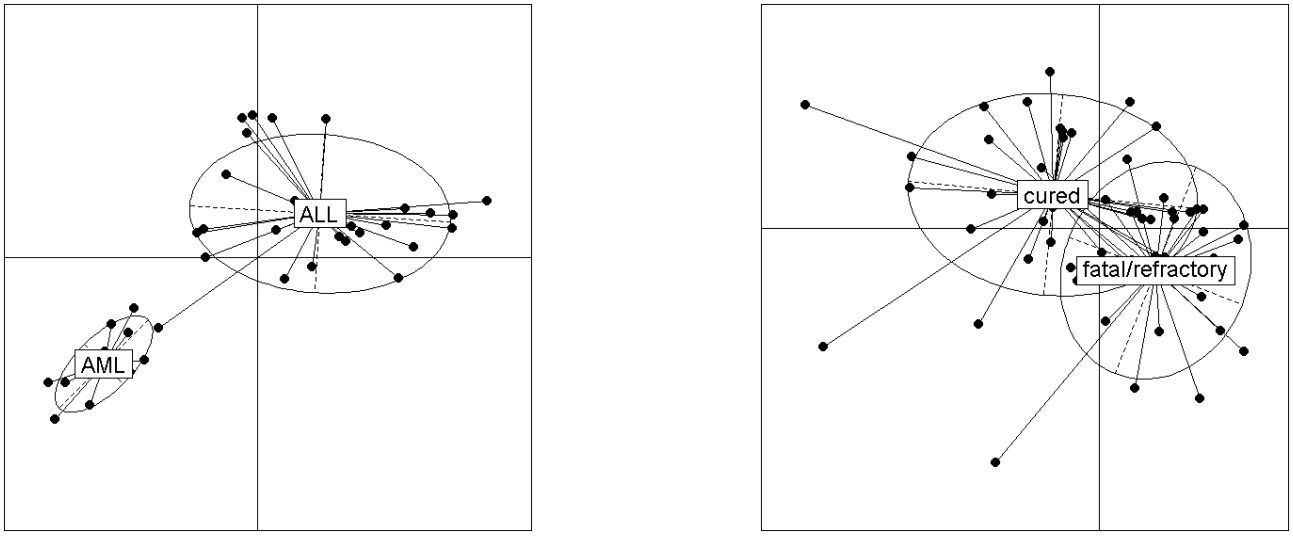


Fig. 3.1: *Représentation des individus dans le premier plan de l'ACP intra-groupes pour les jeux de données de Golub (à gauche) et Shipp (à droite).*

Ces différences portent sur les critères suivants :

Critère 1 : La distance entre les centres de gravité des deux groupes.

Critère 2 : La structure de la matrice de variance/covariance dans chacun des groupes.

Critère 3 : La direction de la droite reliant les barycentres des 2 groupes.

Nous avons souhaité simuler des configurations de structure qui soient intermédiaires entre ces deux jeux de données. Dans un premier temps nous nous sommes intéressés à des modes de simulations proposées dans des articles dont l'objectif était d'évaluer des méthodes d'analyses proches de celles que nous voulions comparer.

Guo *et al.* [100] ont proposé une méthode de simulation pour évaluer les performances d'une méthode d'analyse discriminante régularisée.

Ces simulations permettent d'introduire une structure de dépendance entre les gènes. Les niveaux d'expression de tous les gènes d'un patient (10 000 gènes), c'est à dire son profil génétique, sont issus d'une loi normale multivariée $MVN(0, \Sigma)$. Deux groupes de patients sont ensuite définis, et dans l'un des groupes la valeur 1/2 est ajoutée aux 200 premiers gènes. Biologiquement, cela signifie que certains gènes sont différentiellement exprimés selon le groupe, mais conservent le même système de régulation quel que soit le groupe.

Pour reproduire les réseaux de régulation, la matrice de variance-covariance est scindée en blocs de variables dépendantes ; la structure de covariance dans chacun des b blocs Σ_b est de type auto-régressive, comme l'illustre la figure 3.2. Ce choix permet de représenter des réseaux de régulation en cascade, où l'influence d'un gène sur ses voisins décroît au fur et à mesure de la chaîne de régulation. La valeur absolue de ρ est la même dans tous les blocs ; dans la moitié d'entre eux elle est positive, et dans l'autre négative.

Ce mode de simulation permet de contrôler la structure de variance-covariance dans chacun des groupes (critère 2 ci-dessus), mais il ne permet pas de prendre en compte les deux autres critères ; ce mode de simulation a donc été abandonné.

Nguyen *et al.* [101] ont proposé un mode de simulation pour comparer les réductions de la dimension par ACP ou PLS. Les données d'expression sont construites à partir de $d = 6$ composantes, les trois premières expliquant 33 à 90% de la variance totale. Les niveaux d'expression de chaque individu sont obtenus comme une combinaison linéaire de ces d composantes. Un modèle de régression logistique permet d'assigner un groupe à chaque individu en fonction du niveau d'expression de ses gènes. Ce mode de simulation présente l'intérêt de contrôler le pourcentage

$$\Sigma_b = \begin{pmatrix} 1 & \rho & \cdots & \rho^{29} & \rho^{30} \\ \rho & 1 & \ddots & \ddots & \rho^{29} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{29} & \ddots & \ddots & \ddots & \rho \\ \rho^{29} & \rho^{28} & \cdots & \rho & 1 \end{pmatrix}$$

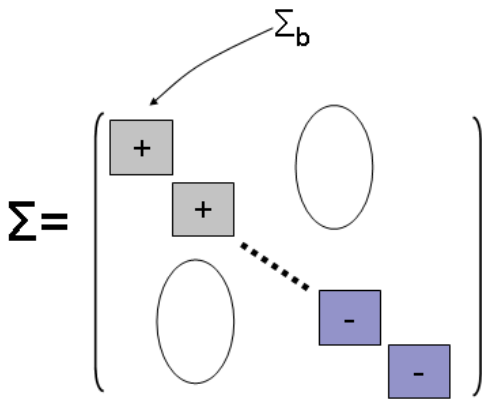


FIG. 3.2: *Structure de la matrice de variance-covariance*

de variance totale expliqué par les composantes. Cependant, il ne permet ni de contrôler la manière dont cette structure se décompose entre variances inter et intra-groupes, ni la structure de la matrice de variance-covariance dans chacun des groupes. Puisqu'il ne permet pas de tenir compte de nos trois critères, nous avons également abandonné ce mode de simulation.

Toujours dans le même contexte, Boulesteix [46] a proposé une méthode de simulation dans le but d'évaluer l'effet du boosting sur les performances de la méthode PLS+DA dans le cas de trois ou quatre classes. Cinquante jeux tests et cinquante jeux de travail contenant 200 gènes pour respectivement 30 et 100 patients par classes, sont générés. Chaque classe k est caractérisée par 10 gènes spécifiques de cette classe et tels que leur niveau d'expression est défini par $x_j/(y = k) \sim N(\mu = 0, \sigma = 1)$ et $x_j/(y \neq k) \sim N(\mu = 1, \sigma = 1)$, $j = 1..p$. Les gènes restants sont issus d'une loi normale centrée réduite dans toutes les classes. Ce mode de simulation permet de contrôler le nombre, l'identité et l'effet des gènes différentiels mais il ne permet pas de modéliser différentes répartitions de la variance totale entre variances inter- et intra-groupes ; il ne répond donc pas non plus à nos attentes.

Finalement, puisque l'espace de l'ACP intra-groupes permet de visualiser les paramètres clé, nous avons choisi de démarrer les simulations dans l'espace défini par les deux premières composantes de l'ACP intra-groupes, en se plaçant dans une situation simple où la différence entre les groupes s'exprime dans ce premier plan.

Nous nous sommes placés dans le cas de deux groupes de $n = 50$ patients chacun, pour lesquels les niveaux d'expression de $p = 500$ gènes sont connus. La matrice de données est de dimension (n, p). Puisque $n << p$, le rang de cette matrice est au maximum n, et par

suite l'espace défini par les composantes de l'ACP intra-groupes est au maximum de dimension n. Les données d'expression ont d'abord été générées dans cet espace des composantes, sous l'hypothèse de variances identiques dans les deux groupes.

Trois paramètres ont permis de contrôler les structures inter et intra-groupes dans ce premier plan factoriel :

1. **Paramètre dist** : Il contrôle la distance entre les deux centres de gravité. Ce paramètre permet de contrôler l'importance de la variance inter-groupes.
2. **Paramètre** α : Ce paramètre permet de contrôler la direction dans laquelle s'exprime la variance inter-groupes, i.e plutôt dans une direction de faible ou grande variance intra-groupes.
3. **Paramètre ratio** : Il contrôle la structure de variance-covariance dans chacun des groupes. Dans l'espace des deux premières composantes, les variables sont issues d'une loi binormale $N(\mu, \Sigma)$. μ est proportionnel au paramètre *dist*, et Σ est une matrice diagonale (2×2) d'éléments σ_1 et σ_2. Ces éléments sont proportionnels aux deux premières valeurs propres de l'ACP intra-groupes. Le ratio σ_1/σ_2 reflète l'excentricité du nuage de points dans l'espace des deux premières composantes : plus il est élevé, plus l'excentricité du nuage est forte. On contrôle ainsi la structure de la variance intra-groupes. D'un point de vue biologique, les composantes peuvent s'interpréter comme des groupes de gènes participant à un même réseau de régulation. Une variance élevée sur une composante traduit un groupe de gènes fortement corrélés. L'excentricité reflète donc d'une certaine manière l'importance des réseaux de régulation.

La figure 3.3 aide à la compréhension géométrique de ces trois paramètres.

On considère que les $n-2$ autres composantes correspondent à du bruit et ne fournissent pas d'information sur la distinction entre les groupes. Les coordonnées des individus sur ces 48 composantes sont issues d'une loi multinormale de moyenne 0 et de matrice de covariance une matrice diagonale de valeur 1. Une rotation permet de passer de l'espace des composantes à un sous-espace des gènes de dimension n. Cette matrice de rotation permet de masquer plus ou moins la structure de variance inter-groupes simulée dans l'espace des deux premières composantes. Nous avons choisi de ne contrôler que les deux premiers axes de cette rotation. Le premier axe est situé dans le plan de la feuille avec un angle β différent de α. Le second axe est dans un plan orthogonal à celui de la feuille, et décrit un angle γ avec le premier axe. Les axes suivants sont construits de telle sorte que la base finale soit orthonormée.

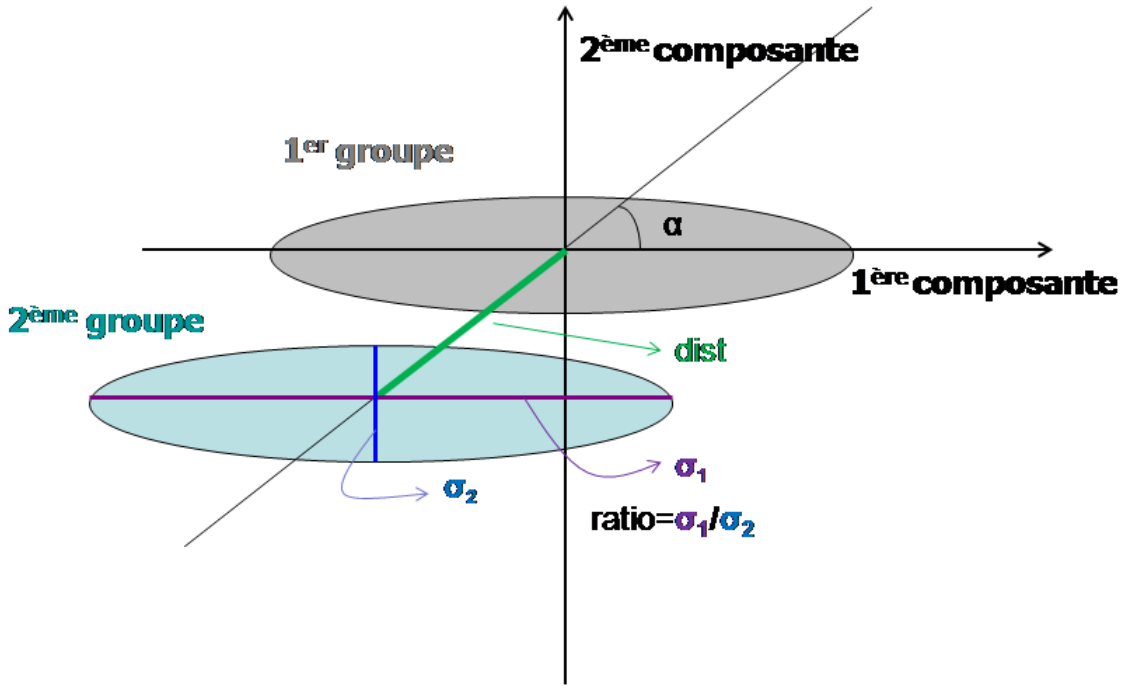


FIG. 3.3: *Visualisation du nuage de points des individus de chacun des groupes dans le premier plan de l'ACP intra-groupes.*

Pour se rapprocher de situations réelles, nous avons fixé $\beta = \pi/4$ et $\gamma = \pi/3$. Les coordonnées des axes de rotation supplémentaires sont issus d'une loi uniforme $U[-1;1]$. La matrice de rotation obtenue est ensuite orthonormalisée.

Les $p - n$ gènes restants sont déterminés par une combinaison linéaire des n premiers. Les coefficients de cette combinaison linéaire sont tirés dans une loi uniforme $U[-1;1]$.

Le schéma de la figure 3.4 reprend les étapes principales des simulations.

Ce mode de simulation a été conçu pour le cas de deux groupes. Le procédé est cependant généralisable à un nombre supérieur de groupes. On pourrait envisager la génération d'un troisième nuage de points, distinct des deux premiers dans l'espace de l'ACP intra-groupes.

3.2.2 Second outil de simulation

Cette fois, l'objectif est de simuler des individus pour lesquels sont disponibles des variables clinico-biologiques classiques et des variables transcritomiques.

On considère une population d'individus pour lesquels on connaît les niveaux d'expression de p gènes, la valeur de deux variables clinico-biologiques classiques, ainsi que le délai de survenue d'un événement quelconque (décès, récidive, etc) et l'indicateur de censure.

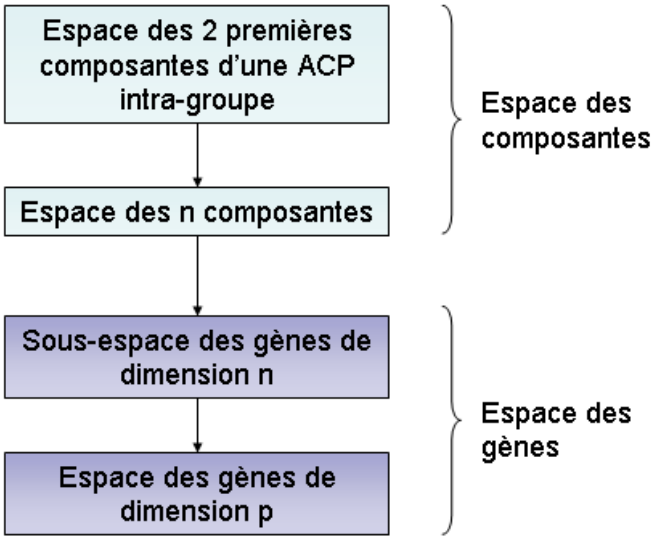


FIG. 3.4: *Etapes principales des simulations*

Les deux variables cliniques sont issues d'une loi binomiale de paramètres 0.5 et 0.4 respectivement.

Chacun des p gènes est issu d'une loi normale centrée réduite $N(0,1)$. Ce choix fait l'hypothèse que les niveaux d'expression des gènes ont été normalisés. Par ailleurs, on considère les gènes comme indépendants, ce qui est une hypothèse simplificatrice mais faite par d'autres auteurs [72, 102]. Parmi les p gènes connus, seuls $p1$ ont un effet sur la survie.

Les $(p+2)$ variables sont ensuite reliées à la survie par l'intermédiaire d'un modèle de Cox. Dans ce modèle, les coefficients des variables cliniques sont fixés à 0.8, ceux des $p1$ gènes à 0.2, et ils sont nuls pour les $p_0 = p - p_1$ gènes restants.

La fonction de base a été simulée par une distribution de Weibull. Les censures ont été générées par une distribution uniforme $U[0,8]$, ce qui a conduit à 40% de données censurées. Pour un ensemble fixé de paramètres p, p_1 et n, 60 jeux de données de travail ont été simulés selon le mode opératoire ci-dessus. A chacun de ces jeux de travail sont associés 50 jeux tests, simulés avec les mêmes paramètres.

Chapitre 4

Importance de la prise en compte de la structure des données dans la comparaison de deux méthodes de réduction de la dimension

4.1 Introduction

Ce travail a fait l'objet d'un article accepté pour publication dans le journal *BMC Bioinformatics* [103]. L'objectif de ce travail a été de comparer les qualités prédictives de deux méthodes évoquées dans la partie 2.3.1 : l'analyse discriminante (AD) précédée d'une ACP ou d'une PLS, et l'analyse inter-groupes. Plus précisément nous avons voulu montrer, en revenant à leurs propriétés théoriques, l'importance de la prise en compte de la structure des données sur leurs performances prédictives.

Dans la littérature relative aux biopuces, ces méthodes de projection ont déjà fait l'objet de comparaisons. Nguyen et Rocke ont comparé les méthodes de réduction par ACP ou PLS dans le cas de deux classes, les composantes obtenues étant introduites comme variables dans des analyses discriminante logistique, linéaire ou quadratique. Cette comparaison était basée sur des jeux de données publics [104] ou simulés [101] suivant le principe décrit dans la partie 3.2.1. Dans leur approche, la réduction de la dimension est précédée d'une sélection d'un nombre restreint de gènes basée sur un test de t. Les auteurs montrent que les performances prédictives

de l'ACP sont moins bonnes que celles de la PLS uniquement quand les gènes sélectionnés dans la première étape sont trop nombreux, ou qu'ils n'ont pas de lien avec le phénotype d'intérêt. Leur critère de performance est le pourcentage de bien classés obtenu par Leave-One-Out Cross-Validation.

Boulesteix [46, 105] a étudié plus en détail la méthode PLS suivie d'une AD (PLS+AD) en se basant d'une part sur neuf jeux de données publics décrivant des phénotypes en deux jusqu'à huit classes, et d'autre part sur des simulations (décrites dans la partie 3.2.1). Le critère de performance est toujours la proportion de bien classés par validation croisée. L'auteur montre que l'utilisation de l'approche PLS+AD donne de meilleures prédictions que les méthodes knn, SVM ou PAM [1]. Elle étudie également l'apport du boosting pour la méthode PLS+AD, qu'elle juge peu intéressant.

Dai *et al.* [106] ont inclus la méthode de réduction de la dimension SIR[2] à la comparaison de l'ACP et de la PLS. Les composantes obtenues sont entrées comme nouvelles variables d'une analyse discriminante logistique. Comme Nguyen *et al.* , ces trois analyses ont été faites après sélection d'un nombre restreint de variables. La comparaison repose sur l'utilisation de deux jeux de données publics sur le colon et la leucémie, tous deux présentés dans la partie 3.1. Les auteurs concluent que les méthodes SIR et PLS sont plus performantes en terme de prédiction que l'ACP, car elles tiennent compte du phénotype des individus dans la réduction de la dimension.

Parallèlement, Jeffery *et al.* [50] ont montré l'influence de la variance intra-groupes du jeu de données sur les performances des méthodes de sélection de gènes. Il nous a semblé intéressant de suivre une démarche similaire en étudiant l'influence de la structure des données sur les méthodes de prédiction citées ci-dessus. Sachant qu'il n'existe pas de méthode idéale quel que soit le jeu de données, deux questions se posent : définir dans quelle situation une méthode est plus adaptée ; et se faire une idée de la structure d'un nouveau jeu de données à analyser pour identifier la méthode a priori la plus adaptée. Une connaissance plus approfondie du rôle de la structure aiderait alors au choix de la méthode la plus adaptée à un nouveau jeu de données.

Pour répondre à cet objectif nous avons utilisé de manière complémentaire des jeux de données publics et simulés. Dans un premier temps les deux méthodes seront présentées, puis les résultats obtenus seront discutés.

[1] cf partie 2.3
[2] cf partie 2.3.2.2

4.2 Matériel et méthodes

4.2.1 Schéma d'analyse général

L'analyse inter-groupes[1] et l'analyse discriminante ont le même objectif : trouver un sous-espace des variables (ici les gènes) dans lequel la variance entre les groupes est maximale. La théorie de ces méthodes a été abordée ici d'un point de vue géométrique dans le cadre de l'analyse multidimensionnelle [107]. Nous avons choisi cette approche géométrique pour mettre en évidence la relation entre ces méthodes et montrer comment elles prennent en compte la structure des données.

Soit un triplet défini par Z, Q et D :

- Z est une matrice (n, p) qui contient p variables pour n individus. Les colonnes de Z sont des vecteurs de $\mathbb{R}^n$; les lignes sont des vecteurs de $\mathbb{R}^p$.
- Q est une matrice définie positive (p, p) qui définit le produit scalaire dans $\mathbb{R}^p$, c'est à dire les distances entre les individus.
- D est une matrice (n, n) qui définit le produit scalaire dans $\mathbb{R}^n$, c'est à dire les distances entre les variables.

Le triplet (Z, Q, D) peut être disposé dans un schéma de dualité [108, 109] :

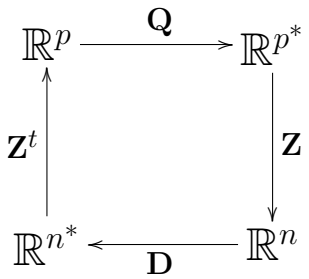


FIG. 4.1: *Schéma de dualité général*

De ce schéma découle un processus unique de "diagonalisation d'un schéma de dualité", qui aboutit à des combinaisons linéaires des variables, $Z\alpha$, qui maximisent $\|Z\alpha\|_D$. Ces combinaisons linéaires définissent un espace dans lequel la variance de Z est maximale. La solution est unique et donnée par la décomposition en valeurs singulières de la matrice QZ^tDZ. Cette matrice est diagonalisable et a p valeurs propres λ_i, $i = 1..p$, parmi lesquelles r sont non nulles, r étant le rang de la matrice Z. Ces r valeurs propres sont positives et telles que $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r$.

[1] Par la suite nous utiliserons l'abréviation BGA, pour Between-Group Analysis

Elles maximisent $\|Z\alpha\|_D$ sous la contrainte de Q^{-1}-orthonormalité. Le premier vecteur propre α_1, associé à λ_1 maximise $\|Z\alpha\|_D$. Le maximum correspondant est λ_1. Le second vecteur propre α_2 maximise $\|Z\alpha\|_D$, et est Q^{-1}-orthogonal à α_1, et ainsi de suite. α est la matrice (p, r) définie par les vecteurs propres en colonne. Ils forment une nouvelle base dans laquelle la variance du nuage de points des individus est maximale.

L'application la plus simple de ce schéma est l'Analyse en Composantes Principales (ACP) [28], dont l'objectif est de définir un sous-espace des variables qui rende maximale la variance totale des données. Dans cette méthode, les individus et les variables ont respectivement le même poids. Cela revient à définir un triplet $(Z, Q, D) = (Z, I_p, \frac{1}{n}(I_n))$, qui conduit au schéma général de la figure 4.2 :

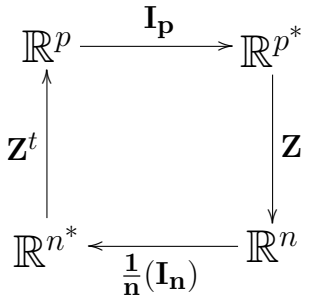


FIG. 4.2: *Schéma de dualité de l'analyse en composantes principales.*

Dans ce cas le plus simple, il n'y a aucune notion sur l'appartenance des individus à des groupes prédéfinis ; c'est la variance totale qui est maximisée. L'objectif de l'AD et de la BGA est de maximiser la variance inter-groupes. Il est donc nécessaire d'introduire de l'information sur les groupes, ce qui conduit à une nouvelle définition du triplet (Z, Q, D).

4.2.2 Choix de Z

Soit X la matrice des données d'expression, avec autant de lignes n que d'individus, et autant de colonnes p que de gènes. Soit Y la matrice (n, k) qui définit la partition des individus en k groupes. Enfin, soit P_Y le projecteur défini par $P_Y = Y(Y^tDY)^{-1}(Y^tD)$. Projeter une variable quelconque sur un vecteur d'indicatrices de classes revient à calculer les moyennes de cette variable dans chacune des classes. $Z = P_Y X$ est une matrice de dimension (n, p) où la valeur de chacune des variables d'un individu est remplacée par la moyenne de cette variable pour le groupe auquel il appartient. Avec ce choix, maximiser la variance de Z revient à maximiser la

variance inter-groupes de X. Les vecteurs α_i, $i = 1, ..., k-1$, correspondant à la décomposition en valeurs singulières de QZ^tDZ sont les axes discriminants ; ils définissent un sous-espace dans lequel les individus sont séparés selon les groupes auxquels ils appartiennent.

4.2.3 Choix de D

Rappelons que D définit le poids des individus pour le calcul des distances entre les variables. Dans le cas des biopuces, la même importance est donnée à tous les individus, ce qui conduit à choisir $D = \frac{1}{n} I_n$.

4.2.4 Choix de Q

Du choix de Q vont résulter deux méthodes différentes : l'analyse discriminante et l'analyse inter-groupes.

4.2.4.1 Analyse inter-groupes

C'est le cas le plus simple, Q étant défini comme la matrice identité (p, p) : $Q = I_p$. Le triplet correspondant est $(Z, I_p, \frac{1}{n} I_n) = (P_Y X, I_p, \frac{1}{n} I_n)$, ce qui correspond aux schémas de la figure 4.3.

Cette analyse correspond à une ACP sur le tableau des moyennes. Dans leur article, Culhane *et al.* [39] proposent une deuxième utilisation de l'analyse inter-groupes, basée sur une Analyse des Correspondances inter-groupes. Dans ce cas, les données d'expression sont vues comme une table de contingence où les gènes et les individus deviennent deux variables qualitatives. Jugeant que les individus ne sont pas réellement des variables qualitatives, nous avons préféré nous concentrer sur la version ACP. Dans leur article, les auteurs ne font pas de préconisation précise quant à l'utilisation privilégiée de l'une ou l'autre forme d'analyse.

4.2.4.2 Analyse discriminante

Cette fois, $Q = \frac{1}{n}(X^tX)^{-1}$, ce qui conduit au schéma de dualité de la figure 4.4.

Les distances entre les individus font intervenir la matrice de variance-covariance de X. Plus concrètement, cela signifie que la structure de variance à l'intérieur de chacun des groupes est prise en compte pour la détermination des axes discriminants, alors qu'elle ne l'est pas dans l'analyse inter-groupes. On peut aussi choisir pour Q la moyenne des variances intra-groupes plutôt que la variance totale. La variance totale se décomposant en variance intra- et inter-

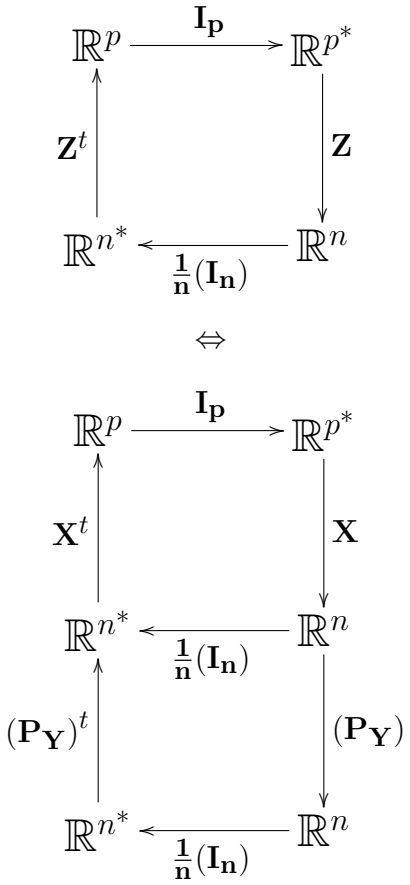
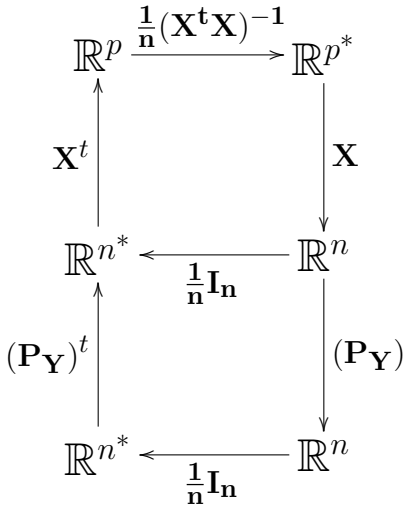


FIG. 4.3: *Schéma de dualité de l'analyse inter-groupes.*



.

FIG. 4.4: *Schéma de dualité de l'analyse discriminante.*

groupes, les deux formes pour Q conduisent aux mêmes valeurs propores à une constante près. Le choix de la variance totale est fait dans l'approche dite "géométrique", tandis que le choix de la variance intra-groupes est fait dans l'approche dite "probabiliste". Dans les deux cas, on fait l'hypothèse que les variances sont les mêmes dans chacun des groupes. Ce choix de Q fait intervenir le calcul de l'inverse de X, qui pose problème dans le cas où $p >> n$, la matrice X étant singulière. Ceci implique pour l'AD une première étape de réduction de la dimension, que nous avons effectuée soit par une ACP (méthode ACP+AD), soit par une approche PLS (méthode PLS+AD).

La différence majeure entre la BGA et l'AD se résume au choix de la métrique Q, qui fait intervenir ou non la structure de variance dans chacun des groupes. C'est donc la structure du jeu de données qui doit être au coeur de la comparaison des performances prédictives de ces méthodes.

4.2.5 Critère de comparaison des méthodes

Nous nous sommes placés dans le cas particulier de deux groupes, qu'un seul axe discriminant suffit à séparer. Sur cet axe est défini le seuil suivant, proposé par Culhane *et al.* [39] :

$$\frac{\bar{X}_{G1} SD_{G2} + \bar{X}_{G2} SD_{G1}}{SD_{G1} + SD_{G2}}$$

où $\bar{X}_{G1}$, $\bar{X}_{G2}$, SD_{G1}, et SD_{G2} sont respectivement les moyennes et écarts-types des coordonnées des individus dans chacun des deux groupes. Ce seuil permet de tenir compte de la variance dans chacun des groupes. Cette pondération n'est pas classique, la pondération la plus usuelle étant l'inverse de la variance.

Pour comparer les performances prédictives des méthodes, nous avons choisi la proportion d'individus bien classés obtenue par validation croisée. A chaque étape de validation croisée, deux tiers des patients sont sélectionnés aléatoirement pour la constitution du jeu de travail, et le tiers restant constitue le jeu test ; ce processus est répété 50 fois. Dans le cas de l'AD, la sélection du nombre optimal de composantes pour l'ACP et la PLS est incluse dans le processus de validation croisée. Celui-ci est répété pour chaque nombre potentiel de composantes et c'est le nombre de composantes qui maximise la proportion de bien classés qui est retenu. Ce même processus avait été employé par Bouleistex [46]. Nous avons contraint le nombre optimal de

composantes retenues à ne pas dépasser 13, après avoir observé que davantage de composantes ne permettait pas d'améliorer les prédictions.

4.3 Résultats

Les résultats présentés ici ont été obtenus sur les jeux de données réels et simulés décrits dans la partie 3.2.1. La figure 4.5 permet de représenter la manière dont les résultats ont été obtenus pour chacun de ces types de données, en bleu pour les jeux de données réels et en rouge pour les jeux de données simulés.

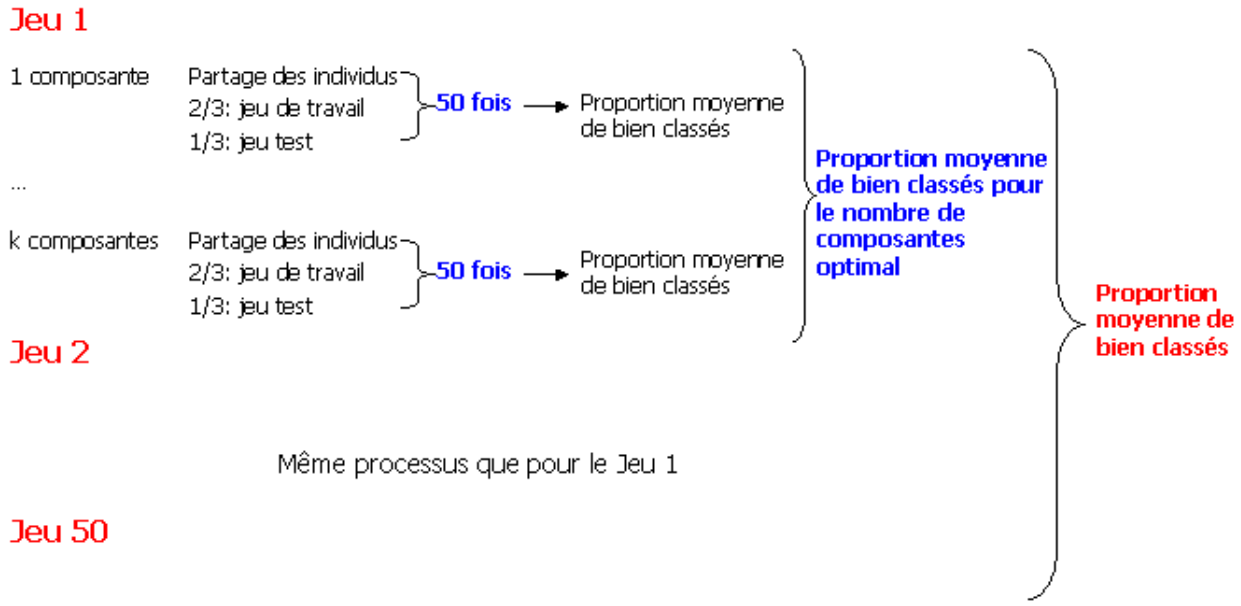


Fig. 4.5: *Mode d'obtention des résultats*

A chaque proportion de bien classés a été associé un écart-type. Son calcul est différent selon que le jeu de données est réel ou simulé. Dans le premier cas, il correspond à l'écart-type des proportions de bien classés sur les 50 étapes de validation croisée pour le nombre optimal de composantes (en bleu sur la figure 4.5). Dans le second cas, il correspond à l'écart-type obtenu sur l'ensemble des 50 jeux de données générés pour chaque jeu de paramètres (en rouge sur la figure 4.5).

	$\alpha = \pi/2$	$\alpha = \pi/3$	$\alpha = \pi/4$	$\alpha = \pi/6$	$\alpha = 0$
$dist = 1$					
PLS+AD	0.69(0.05) [2]	0.69(0.06) [2]	0.64(0.06) [2]	0.60(0.06) [2]	0.59(0.05) [1]
ACP+AD	0.69(0.04) [3]	0.70(0.05) [2]	0.66(0.06) [3]	0.60(0.05) [3]	0.58(0.06) [1]
BGA	0.63(0.05)	0.61(0.06)	0.55(0.06)	0.57(0.05)	0.58(0.05)
$dist = 3$					
PLS+AD	0.93(0.04) [2]	0.91(0.03) [2]	0.86(0.03) [2]	0.71(0.03) [2]	0.69(0.06) [1]
ACP+AD	0.94(0.04) [2]	0.91(0.03) [2]	0.85(0.04) [3]	0.71(0.04) [3]	0.69(0.05) [2]
BGA	0.90(0.04)	0.79(0.03)	0.73(0.03)	0.70(0.03)	0.67(0.06)
$dist = 5$					
PLS+AD	0.98(0.04) [2]	0.97(0.01) [2]	0.97(0.02) [2]	0.84(0.03) [2]	0.79(0.04) [1]
ACP+AD	0.99(0.01) [3]	0.98(0.01) [2]	0.97(0.02) [2]	0.83(0.03) [2]	0.79(0.04) [2]
BGA	0.91(0.04)	0.91(0.01)	0.86(0.03)	0.82(0.03)	0.79(0.04)

Tab. 4.1: *Effet de la distance dist sur la proportion de bien classés - Moyenne (écart-type) [médiane du nombre optimal de composantes], estimés sur 50 jeux de données simulés avec ratio = 10.*

4.3.1 Jeux de données simulés

Rappelons que pour générer des structures de données différentes, trois paramètres ont été utilisés : 1-le paramètre *ratio*, qui intervient sur la structure de variance-covariance dans chacun des groupes ; 2- le paramètre *dist*, qui définit la distance entre les centres de gravité des deux groupes dans le premier plan de l'ACP intra-groupes ; 3- le paramètre α, qui indique dans quelle direction s'exprime cette distance dans ce même plan.

Pour chaque jeu de paramètres sont simulés 50 jeux de données. Les résultats suivants montrent l'influence de chacun de ces paramètres.

4.3.1.1 Effet de la distance entre les barycentres

Le tableau 4.1 permet l'étude de l'influence de la distance *dist* entre les barycentres sur les performances prédictives des méthodes.

Quel que soit la valeur des paramètres, la proportion d'individus bien classés est supérieure pour l'AD, et ceci quel que soit la méthode de réduction de dimension préalablement employée (PLS ou ACP). Les résultats obtenus pour l'ACP et la PLS sont très proches et le nombre de composantes retenu par l'ACP est toujours supérieur ou égal à celui retenu par la PLS.

Quelles que soient la valeur de α et la méthode utilisée, la prédiction s'améliore quand les nuages de points s'éloignent.

Quand α se rapproche de zéro, la prédiction diminue. L'influence de l'angle n'est pas la même pour les deux méthodes. C'est avec $\alpha = 0$ et $\alpha = \pi/2$ que les deux méthodes sont les plus proches. Ces cas correspondent respectivement à la simulation de la variance inter-groupes le long de la première et de la deuxième composante de la variance intra-groupes. Dans le premier cas les résultats sont mauvais alors que c'est dans le second qu'ils sont les meilleurs. Avec un angle $\alpha = \pi/4$ la supériorité de l'AD par rapport à la BGA est la plus forte.

4.3.1.2 Effet de l'excentricité

L'influence de la forme des nuages de points a été étudiée par l'intermédiaire du paramètre d'excentricité *ratio*. Les résultats figurent dans le tableau 4.2.

Un ratio de 1 correspond à un nuage de points sphérique. Plus les nuages de points sont excentrés, ce qui correspond à un ratio qui s'éloigne de 1, plus les performances prédictives de l'AD surpassent celles de la BGA. A part pour $\alpha = 0$, les méthodes se comportent mieux quand le ratio est élevé, ce qui correspond à une faible excentricité. Quand le ratio diminue, les performances des méthodes se rapprochent.

Dans le cas où les nuages de points sont sphériques, les méthodes se comportent de la même manière et subissent peu l'influence de la valeur de l'angle α.

Concernant l'angle α, les observations faites ci-dessus restent valables : plus cet angle se rapproche de 0, moins les méthodes sont performantes. C'est avec $\alpha = \pi/4$ que la différence entre les deux méthodes est la plus marquée.

Nous avons également évalué les performances prédictives des méthodes dans le cas où les directions principales de variance dans chacun des groupes étaient différentes (Tableau 4.3). Dans ce cas, PLS+AD et BGA ont des résultats équivalents, tandis que la méthode d'ACP+AD est moins performante.

4.3.1.3 Interprétation des résultats

D'un point de vue théorique, la différence majeure entre la BGA et l'AD se résume au choix de la métrique Q, qui fait intervenir ou non la structure de variance dans chacun des groupes. La BGA projette orthogonalement les individus sur l'axe discriminant, tandis que l'AD projette les individus dans la direction principale de la variance totale. A travers les trois paramètres *dist*, α et *ratio*, les simulations nous ont permis d'identifier trois configurations particulières de la structure des données.

	$\alpha = \pi/2$	$\alpha = \pi/3$	$\alpha = \pi/4$	$\alpha = \pi/6$	$\alpha = 0$
$ratio = 10$					
PLS+AD	0.82(0.05) [2]	0.81(0.03) [2]	0.76(0.03) [2]	0.71(0.04) [2]	0.59(0.04) [2]
ACP+AD	0.85(0.05) [3]	0.81(0.04) [3]	0.77(0.05) [3]	0.73(0.05) [3]	0.59(0.04) [3]
BGA	0.76(0.05)	0.75(0.04)	0.66(0.03)	0.67(0.04)	0.58(0.04)
$ratio = 2$					
PLS+AD	0.68(0.05) [1]	0.65(0.04) [1]	0.65(0.05) [1]	0.65(0.05) [2]	0.63(0.05) [1]
ACP+AD	0.69(0.05) [3]	0.65(0.04) [3]	0.67(0.04) [2]	0.65(0.04) [2]	0.63(0.04) [2]
BGA	0.67(0.06)	0.62(0.04)	0.64(0.05)	0.65(.04)	0.62(0.05)
$ratio = 1$					
PLS+AD	0.60(0.05) [2]	0.62(0.05) [1]	0.64(0.05) [2]	0.62(0.05) [1]	0.61(0.05) [1]
ACP+AD	0.63(0.04) [3]	0.63(0.04) [2]	0.63(0.05) [2]	0.64(0.05) [2]	0.63(0.05) [2]
BGA	0.61(0.05)	0.62(0.05)	0.61(0.05)	0.61(0.05)	0.60(0.05)

Tab. 4.2: *Effet de l'excentricité (ratio) sur la proportion de bien classés- Moyenne (écart-type) [médiane du nombre optimal de composantes], estimés sur 50 jeux de données simulés avec dist* $= 2$.

	$\alpha = \pi/2$	$\alpha = \pi/3$	$\alpha = \pi/4$	$\alpha = \pi/6$	$\alpha = 0$
PLS+AD	0.70(0.04) [1]	0.70(0.05) [1]	0.70(0.05) [1]	0.70(0.04) [1]	0.70(0.06) [1]
ACP+AD	0.61(0.04) [2]	0.62(0.05) [2]	0.62(0.05) [1]	0.61(0.04) [1]	0.58(0.05) [2]
BGA	0.71(0.04)	0.69(0.05)	0.69(0.06)	0.70(0.04)	0.71(0.06)

Tab. 4.3: *Effet de variances différentes dans chacun des groupes sur la proportion de bien classés- Moyenne (écart-type) [médiane du nombre optimal de composantes], estimés sur 50 jeux de données simulés avec dist* $= 2$.

1. Les nuages de points sont éloignés. Dans ce cas, les deux méthodes donnent des résultats similaires car le mode de projection sur l'axe discriminant n'intervient pas.
2. Les nuages de points sont intriqués ou ont des formes très différentes. Dans ce cas, les deux méthodes sont toutes deux inefficaces. Si les variances sont différentes dans chacun des groupes, la variance intra-groupes ne reflète pas la variance dans chacun des groupes, et projeter les nuages de points dans cette direction ne permet pas de tenir compte de la forme de chacun des nuages. L'absence de variance commune aux deux groupes ne permet pas à l'AD de tirer bénéfice de la prise en compte de la structure des données.
3. Dans des situations intermédiaires, l'AD donne de meilleurs résultats que la BGA car elle permet de tenir compte de la structure de variance particulière des données.

Les observations sur les simulations peuvent alors servir de guide à l'utilisation de ces méthodes dans le cas de jeux de données réels. Pour cela, il est nécessaire de repérer sur ces jeux de données les paramètres clés repérés sur les simulations.

Nous avons donc proposé un outil qui permette de visualiser la structure du jeu de données et plus particulièrement la répartition de la variance totale entre variances inter- et intra-groupes, ainsi que la position relative des nuages de points dans chacun des groupes. A cet effet, nous avons choisi de représenter simultanément les coordonnées des individus dans l'espace défini par une analyse inter-groupes (en abcisse) et celui défini par une analyse intra-groupes (en ordonnée). Deux graphiques sont proposés selon que les coordonnées de l'analyse intra-groupes sont ceux obtenus sur la première ou la deuxième composante. La variance inter-groupes est essentiellement due aux gènes différentiels, tandis que les autres gènes masquent la structure inter-groupes. Pour mettre en évidence cette structure, nous avons choisi de ne considérer que les r gènes les plus différentiels d'après un test de t, r étant le rang de la matrice de données. Dans la suite de ce document nous appellerons ces graphiques les graphiques "inter-intra".

4.3.2 Jeux de données publics

Les jeux de données ont été choisis pour couvrir les principales configurations rencontrées dans la pratique. Les figures 4.6 à 4.15 montrent les graphiques "inter-intra" obtenus pour chacun d'eux. Ces graphiques sont interprétés en parallèle avec les résultats du tableau 4.4.

On peut regrouper les jeux de données en trois groupes selon les trois configurations de la structure identifiées par les simulations auxquelles ils appartiennent :

	PLS+AD	ACP+AD	BGA
DLBCL.1	0.51(0.14) [12]	0.49(0.09) [13]	0.43(0.10)
DLBCL.2	0.97(0.03) [3]	0.96(0.03) [10]	0.84(0.08)
Prostate	0.97(0.06) [10]	0.96(0.07) [9]	0.70(0.09)
Colon	0.87(0.06) [2]	0.83(0.06) [5]	0.88(0.06)
Myélome	0.79(0.10) [1]	0.72(0.05) [12]	0.78(0.04)
ALL.1	0.99(0.01) [2]	0.99(0.01) [5]	0.99(0.01)
ALL.2	0.73(0.05) [10]	0.57(0.08) [1]	0.60(0.06)
ALL.3	0.57(0.07) [6]	0.59(0.08) [1]	0.52(0.07)
ALL.4	0.82(0.07) [4]	0.59(0.08) [6]	0.73(0.09)
Leucémie	0.97(0.03) [1]	0.95(0.04) [5]	0.98(0.03)

TAB. 4.4: *Proportion de bien classés pour les jeux publics - Moyenne (écart-type) obtenus avec le nombre optimal de composantes (entre crochets) sur les 50 étapes de validation croisée correspondantes.*

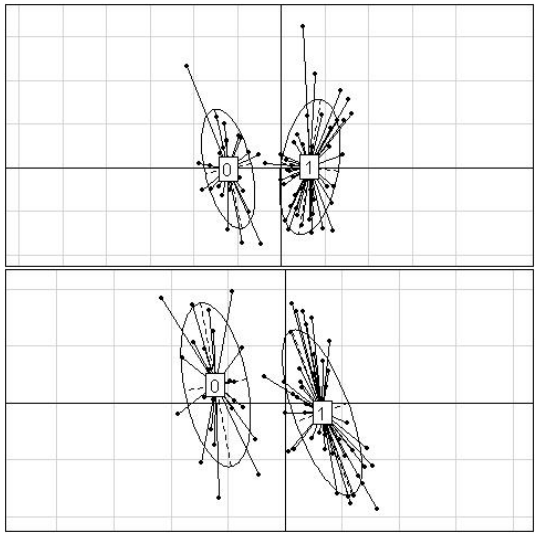


FIG. 4.6: *Visualisation inter-intra du jeu de données Leucémie - 0 : AML ; 1 : ALL. En abcisse figurent les coordonnées des individus sur l'axe de l'analyse inter-groupes, et en ordonnées les coordonnées des individus sur la première (en haut) et la seconde (en bas) composantes de l'analyse intra-groupes.*

1. **Les nuages de points sont distincts**

   C'est le cas des jeux de données Leucémie (Figure 4.6), ALL.1 (Figure 4.7), Colon (Figure 4.8) et Myélome (Figure 4.9). Les deux premiers jeux de données sont particulièrement caricaturaux et permettent un très bon classement des individus. Sur la figure 4.6, les nuages de points sont séparés essentiellement le long de la direction de variance inter-groupes. Cela correspond aux cas simulés avec $\alpha = \pi/2$, qui donnent les meilleures prédictions. Sur la figure 4.7, les nuages de points se distinguent en plus le long de la première composante de l'ACP intra-groupes. Cela correspond aux cas simulés avec α entre 0 et $\pi/2$,

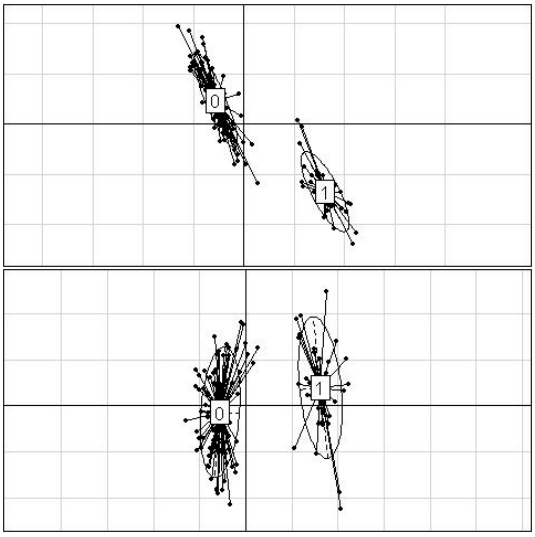


Fig. 4.7: *Visualisation inter-intra du jeu de données ALL.1 - 0 : Origine B-Cellulaire; 1 : Origine T-Cellulaire. En abcisse figurent les coordonnées des individus sur l'axe de l'analyse inter-groupes, et en ordonnées les coordonnées des individus sur la première (en haut) et la seconde (en bas) composantes de l'analyse intra-groupes.*

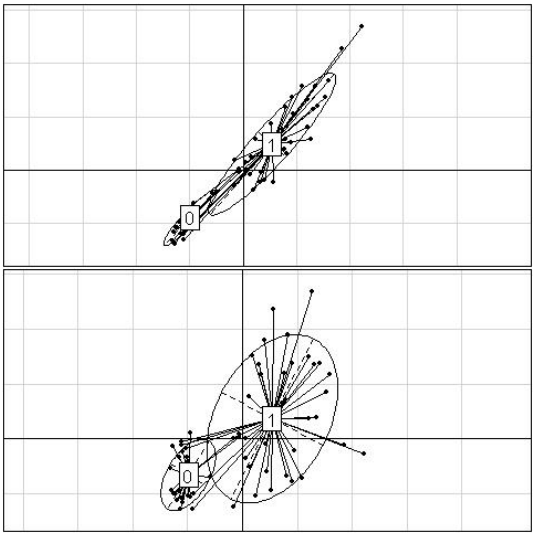


Fig. 4.8: *Visualisation inter-intra du jeu de données Colon - 0 : Échantillon non tumoral; 1 : Échantillon tumoral. En abcisse figurent les coordonnées des individus sur l'axe de l'analyse inter-groupes, et en ordonnées les coordonnées des individus sur la première (en haut) et la seconde (en bas) composantes de l'analyse intra-groupes.*

et une distance assez importante pour que les nuages de points ne se chevauchent pas. La même observation est faite sur la figure 4.9, mais dans ce dernier cas, les nuages sont moins nettement séparés, ce qui conduit à une proportion d'individus bien classés moins importante que pour les deux premiers jeux de données. Enfin, dans le cas du jeu Colon, les groupes sont distingués dans les deux premières directions principales de la variance intra-groupes, ce qui correspond aux cas simulés avec $\alpha = \pi/4$. D'après les résultats de

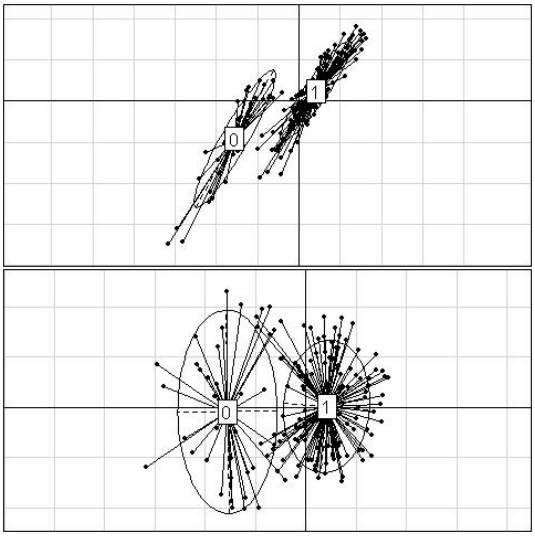


FIG. 4.9: *Visualisation inter-intra du jeu de données Myélome - 0 : Présence ; 1 : Absence d'une région lytique. En abcisse figurent les coordonnées des individus sur l'axe de l'analyse inter-groupes, et en ordonnées les coordonnées des individus sur la première (en haut) et la seconde (en bas) composantes de l'analyse intra-groupes.*

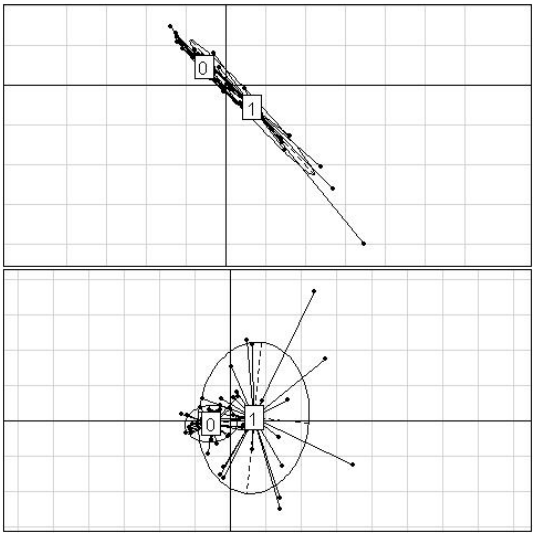


FIG. 4.10: *Visualisation inter-intra du jeu de données DLBCL.1 - 0 : Guérison ; 1 : Rechute. En abcisse figurent les coordonnées des individus sur l'axe de l'analyse inter-groupes, et en ordonnées les coordonnées des individus sur la première (en haut) et la seconde (en bas) composantes de l'analyse intra-groupes.*

simulations, on s'attend à ce que les méthodes soient moins performantes. Les résultats du tableau 4.4 confirment ces observations.

2. **Les nuages de points sont peu distincts**

   C'est le cas des jeux de données DLBCL.1 (Figure 4.10) et ALL.3 (Figure 4.11), où les nuages de points sont intriqués l'un dans l'autre. De plus, la structure de variance-covariance est différente dans chacun des groupes. On s'attend donc à ce qu'aucune des

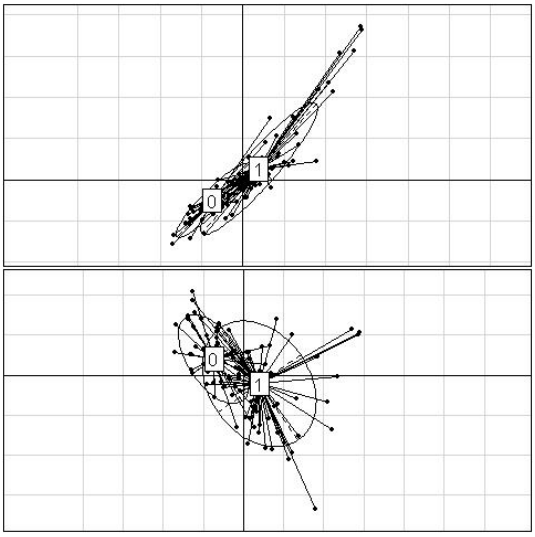


Fig. 4.11: *Visualisation inter-intra du jeu de données ALL.3 - 0 : Guérison; 1 : Rechute. En abcisse figurent les coordonnées des individus sur l'axe de l'analyse inter-groupes, et en ordonnées les coordonnées des individus sur la première (en haut) et la seconde (en bas) composantes de l'analyse intra-groupes.*

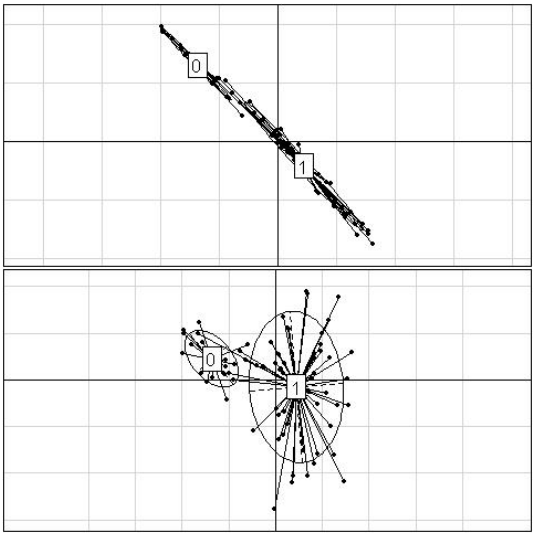


Fig. 4.12: *Visualisation inter-intra du jeu de données DLBCL.2 - 0 : Folliculaire; 1 : Germinal. En abcisse figurent les coordonnées des individus sur l'axe de l'analyse inter-groupes, et en ordonnées les coordonnées des individus sur la première (en haut) et la seconde (en bas) composantes de l'analyse intra-groupes.*

deux méthodes ne soit performante. C'est bien ce qui est observé dans le tableau 4.4. Dans le cas de DLBCL.1, le nombre de composantes retenu pour la PLS (12) et l'ACP (13) est élevé, ce qui permet à l'AD un léger bénéfice par rapport à la BGA. La structure de variance n'étant pas la même dans chacun des groupes, ce bénéfice ne peut être plus important.

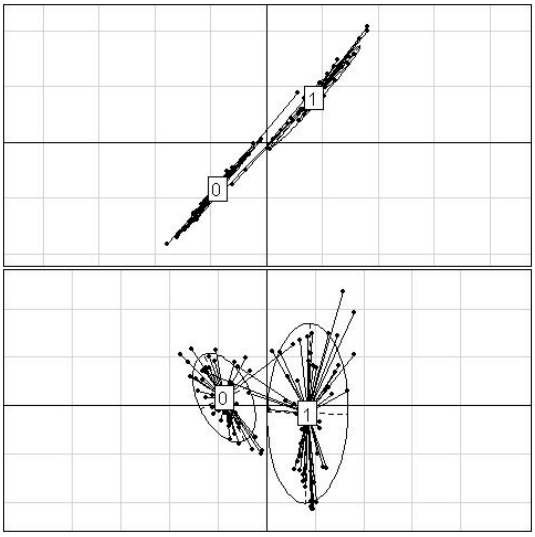


FIG. 4.13: *Visualisation inter-intra du jeu de données Prostate - 0 : Non porteur; 1 : Porteur d'une tumeur. En abcisse figurent les coordonnées des individus sur l'axe de l'analyse inter-groupes, et en ordonnées les coordonnées des individus sur la première (en haut) et la seconde (en bas) composantes de l'analyse intra-groupes.*

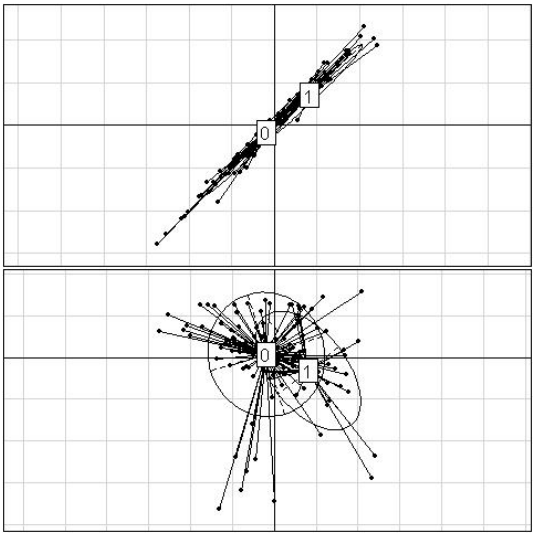


FIG. 4.14: *Visualisation inter-intra du jeu de données ALL.2 - 0 : Multirésistance aux médicaments; 1 : Pas de multirésistance aux médicaments. En abcisse figurent les coordonnées des individus sur l'axe de l'analyse inter-groupes, et en ordonnées les coordonnées des individus sur la première (en haut) et la seconde (en bas) composantes de l'analyse intra-groupes.*

3. **Les nuages de points sont dans une situation intermédiaire**

   C'est le cas des jeux de données DLBCL.2 (Figure 4.12), Prostate (Figure 4.13), ALL.2 (Figure 4.14), et ALL.4 (Figure 4.15).

   Pour le premier jeu de données, les groupes sont distingués dans les directions de la variance inter et intra-groupes. Ceci correspond à une situation simulée où α est entre 0

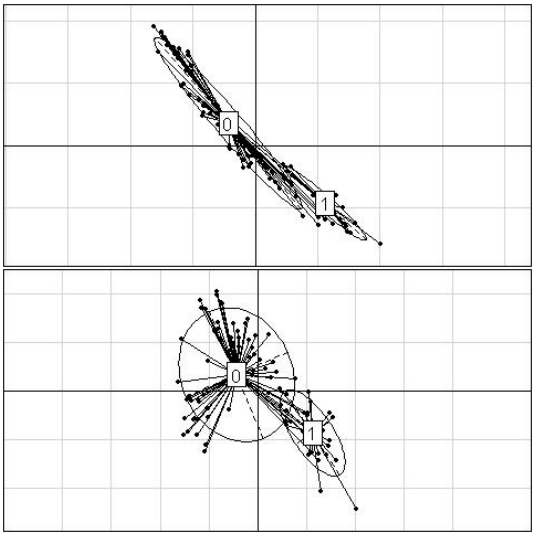


Fig. 4.15: *Visualisation inter-intra du jeu de données ALL.4 - 0 : Absence ; 1 : Présence de la translocation. En abcisse figurent les coordonnées des individus sur l'axe de l'analyse inter-groupes, et en ordonnées les coordonnées des individus sur la première (en haut) et la seconde (en bas) composantes de l'analyse intra-groupes.*

et $\pi/2$. Les mêmes remarques sont valables pour le jeu ALL.4. On s'attend donc à ce que l'AD soit plus performante que la BGA. C'est bien ce qui est confirmé dans le tableau 4.4. Notons que pour les deux jeux ALL, les résultats avec l'ACP sont nettement moins bons que ceux obtenus avec les méthodes BGA et PLS+AD. Ce sont les seuls cas (avec le Myélome dans une moindre mesure), où les réductions préalables PLS et ACP conduisent à des résultats qui ne vont pas dans le même sens.

4.3.3 Remarques sur la taille de l'échantillon

Comme cela a été montré, la méthode d'analyse discriminante prend en compte la structure de variance-covariance du jeu de données analysé lors de la construction de l'axe discriminant. Puisque la particularité de l'analyse des biopuces est que le nombre de patients est faible relativement au nombre de gènes étudiés, la structure de variance-covariance intra-groupes est mal estimée. On pourrait donc présager de moins bonnes performances pour l'analyse discriminante dans le cas où les effectifs sont trop faibles. Cependant dans ce travail, l'analyse discriminante n'a pas été effectuée dans l'espace d'origine des gènes, mais dans le sous espace de projection obtenu par une ACP ou une PLS. Dans cet espace de dimension réduit, la structure de covariance peut être estimée de manière efficace. Ainsi, l'AD n'est pas davantage pénalisée par la taille de l'échantillon que la BGA.

Pour des critères de simulation comparables, les résultats énoncés ci-dessus sur la comparaison entre l'AD et la BGA restent valables quelque soit le nombre d'individus inclus dans l'étude. On peut cependant noter qu'en augmentant la taille de l'échantillon, l'écart-type de la proportion d'individus bien classés diminue.

4.3.4 Remarques sur le choix du nombre de composantes

L'objectif de ce travail était d'évaluer les capacités prédictives de l'ACP et de l'AD. De ce fait, nous n'avons pas cherché dans les simulations à générer des situations qui différencient PLS et ACP et les nombre de composantes retenues pour l'ACP et la PLS sont proches. Les jeux de données réels permettent d'apporter certaines remarques complémentaires à ce sujet.

Pour commencer, ils ont permis de retrouver les conclusions d'autres auteurs selon lesquelles l'ACP est globalement moins adaptée que la PLS. C'est ce qui s'observe sur tous les jeux de données, hormis ceux pour lesquels les nuages de points correspondant aux deux groupes sont nettement distincts.

Les résultats obtenus permettent également de mettre en évidence une propriété montrée initialement par Barker et Rayens [110] et reprise par Boulesteix [46]. Ils ont montré que dans le cas de deux groupes, la première composante obtenue par PLS est identique à l'axe discriminant de la BGA. L'axe discriminant de l'AD qui intègre la première composante PLS comme variable est donc colinéaire à celui de la BGA. Cette propriété est mise en évidence quand les nuages de points sont nettement distincts, où une seule composante PLS est suffisante et mène aux mêmes résultats que la BGA. C'est ce qui est observée pour les jeux de données sur la leucémie ou le myélome. Dans les cas des jeux de données ALL.3 et Colon, une deuxième composante est requise, mais sans que les résultats ne soient améliorés par rapport à la BGA.

Cette observation conduit à une remarque quant à la stabilité du nombre optimal de composantes. En soumettant plusieurs fois un jeu de données aux méthodes PLS+AD ou ACP+AD, nous avons observé que le nombre de composantes optimal varie. Les graphiques de la figure 4.16 montrent l'évolution de la proportion de bien classés en fonction du nombre de composantes retenu pour l'ACP et la PLS pour le jeu de données Colon. Attention, les échelles ne sont pas les mêmes pour les deux graphiques. Les méthodes PLS+AD et ACP+AD ont été appliquées trois fois à ce jeu de données, chacune représentée par une autre couleur. D'une fois sur l'autre, les seules différences se situent au niveau de la manière dont le jeu de données est scindé entre jeu de travail et jeu test lors des étapes de validation croisée. Les variations peuvent donc s'ex-

pliquer par des fluctuations d'échantillonnage dues aux faibles effectifs de patients. Pour l'ACP, la proportion de bien classés atteint un palier à partir de cinq composantes. L'ajout de davantage de composantes améliore peu la proportion de bien classés, et fluctue faiblement. Pour la PLS, les fluctuations sont également faibles relativement à l'échelle de la figure. Le nombre de composantes n'est donc pas une vérité absolue puisque pour des nombres de composantes différents on a parfois des performances très proches. Dans certains cas la proportion de bien classés est très peu modifiée par l'ajout ou la suppression d'une composante.

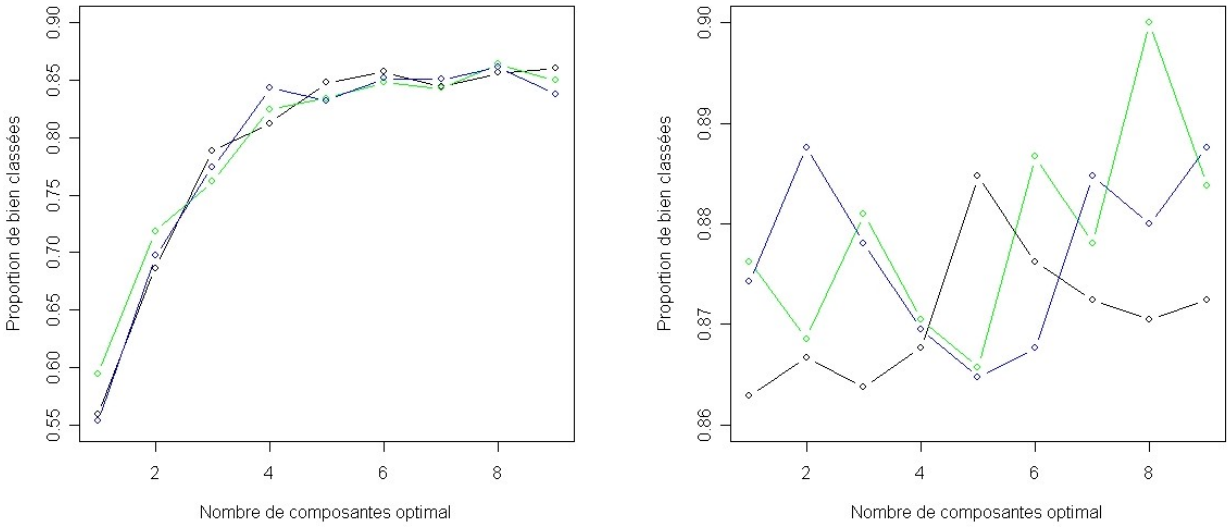


FIG. 4.16: *Évolution de la proportion de bien classés en fonction du nombre de composantes retenu par ACP (à gauche) ou PLS (à droite) pour le jeu de données Colon.*

4.4 Discussion et conclusion

Les résultats obtenus sur les deux types de données, réelles ou simulées, ont montré l'importance de la structure du jeu de données dans l'évaluation des qualités prédictives des méthodes. Pour cela, l'utilisation de jeux de données réels et simulés est complémentaire. Pour tester une nouvelle méthode, la simulation de données qui ont une structure connue permet d'étudier l'influence de la structure sur les performances d'une méthode donnée. Les observations formulées sur les simulations peuvent ensuite servir à guider l'interprétation des résultats sur des jeux de données réels. Face à un nouveau jeu de données à analyser, nous pensons qu'il est indispensable de s'intéresser dans un premier temps à sa structure. Ce sont les caractéristiques de cette structure qui vont orienter le choix de la méthode la plus adaptée. Dans cet objectif, le mode de simulation que nous avons proposé permet de considérer trois critères majeurs : 1-la

structure de variance-covariance dans chacun des groupes, grâce au paramètre *ratio* ; 2-la distance entre les centres de gravité des deux groupes, grâce au paramètre *dist* ; 3-la direction de cette différence grâce au paramètre α. A travers ces critères, notre étude a permis l'identification de trois configurations. Quand les nuages de points sont clairement distincts, la distinction entre les groupes est telle que la structure de variance intra-groupes n'intervient pas dans la projection des individus sur l'axe discriminant. A l'opposé, il existe des situations où aucune des deux méthodes n'est adaptée. C'est le cas lorsque les nuages de points se superposent ou lorsque la variance n'est pas la même dans chacun des groupes. Dans les autres situations, l'AD permettra de meilleures prédictions car elle prend en compte la structure de variance de chacun des groupes. L'utilisation de l'AD semble donc préférable à celle de la BGA car elle permet de s'adapter à des structures de données plus diverses et complexes.

L'utilisation seule de jeux de données nous semble insuffisante. En effet, cette utilisation doit être considérée avec prudence car les jeux de données peuvent ne pas être adaptés, comme c'est le cas du jeu de données de Golub, très largement utilisé (148 citations selon CiteSeer). La visualisation graphique de ces données a mis en évidence le fait que sa structure est particulièrement caricaturale (figure 4.6). Les nuages de points sont tels que n'importe quelle méthode serait capable de les séparer. Ce jeu de données ne permet donc pas d'évaluer les performances prédictives d'une nouvelle méthode. Si, dans un premier temps, sa structure avait été analysée, il n'aurait pas servi de référence pour valider les capacités prédictives de nombreuses méthodes. Nous tenons donc à mettre en garde contre l'utilisation mal appropriée de jeux de données pour lesquels la structure n'a pas été étudiée au préalable.

Nous pensons par ailleurs que la structure d'un jeu de données dépend en partie de la nature du phénotype étudié, selon qu'il relève de facteurs pronostiques ou diagnostiques. Dans le cadre diagnostique, la discrimination entre les classes peut reposer d'un point de vue biologique sur des entités physiopathologiques. Dans le cas où les deux classes ont pour origine des voies d'activation métabolique différentes par exemple, des processus cellulaires différents sont mis en oeuvre dans chacune des classes ; par suite, ce ne sont pas les mêmes gènes qui seront sollicités dans chacune de ces classes et celles-ci seront donc facilement distinguables. Dans le cas où cette distinction ne relève pas d'entités physiopathologiques, la séparation est plus difficile. Ceci est illustré par exemple par le jeu de données ALL.2, où les patients présentent ou non une multirésistance aux médicaments. Cette fois, les classes sont difficilement séparables, et

ceci quel que soit la méthode utilisée. Dans le cas des études pronostiques, l'objectif est le plus souvent de séparer des patients de bon ou mauvais pronostic pour des patients qui ont une même maladie et partagent donc des caractéristiques physiopathologiques communes. Dans ce cas, la séparation des classes est moins évidente.

Dans ce travail nous nous sommes intéressés à des méthodes de discrimination linéaires. Les données ont été simulées selon des hypothèses fidèles à celle d'une analyse discriminante linéaire. La comparaison des méthodes étant restreinte à trois variantes d'analyse discriminante linéaire, ce mode de simulation n'a ici pas de conséquences sur les résultats. Il serait intéressant de reproduire un travail similaire pour d'autres méthodes de discrimination que celles basées sur l'analyse discriminante linéaire, en incluant des cas où la frontière qui sépare les classes est non linéaire. Pour cela, le mode de simulation devrait être adapté, afin que la configuration des simulations n'aient pas de conséquence sur les résultats.

Chapitre 5

Approche comparative de l'optimisme dans les modèles intégrant des variables clinico-biologiques classiques et des gènes

5.1 Introduction

L'introduction des biopuces dans le domaine clinique a donné grand espoir aux cliniciens d'améliorer la compréhension des mécanismes cancéreux, et de découvrir de nouveaux biomarqueurs permettant d'améliorer la prise en charge thérapeutique de leurs patients. Avec l'engouement des premières études, certains auteurs ont affirmé que les biomarqueurs issus de l'étude du transcriptome avaient de meilleures capacités prédictives que les biomarqueurs clinico-biologiques connus jusqu'à maintenant. Ainsi, Shipp *et al.* [8] ont montré qu'une signature de 13 gènes permettait d'affiner la distinction entre patients de bons ou mauvais pronostic fournie par l'IPI (International Prognostic Index)[1]. En 2002, Tibshirani et Efron [34] mettent cependant en garde contre des conclusions trop hâtives et suggèrent que l'information issue des puces à ADN n'est pas aussi forte que celle issue de variables clinico-biologiques classiques. Ils posent pour la première fois la question de la comparaison des prédicteurs cliniques et transcriptomiques. Leurs propos sont illustrés par l'étude de Van't Veer *et al.* [9] sur le cancer du sein, qui identifie une signature de 70 gènes prédictifs de la survenue de métastases à cinq ans. Ils montrent que l'effet des gènes est en réalité surestimé par rapport à celui des variables cliniques

[1] L'IPI est un outil clinique développé par les oncologues pour aider à prédire le pronostic de patients atteints d'un lymphome non Hodgkinien agressif. Cet index intègre les critères suivants : l'âge (>60 ans ou non), le statut de la maladie, le nombre de sites extra-nodaux, le taux de LDH, le niveau d'état de santé général.

classiques telles que le grade ou la taille de la tumeur. Ceci est dû au fait que le modèle incluant les deux types de variables est évalué sur le jeu de données ayant permis de sélectionner les gènes. Les auteurs proposent une méthode de "pré-validation" basée sur la validation croisée pour corriger l'optimisme relié au transcriptome en l'absence de jeu de validation externe.

Dans la continuité de cette réflexion, nous pensons que la situation pour les deux types de variables est totalement différente ; cela doit être pris en compte lors de la construction d'un modèle prédictif alliant les deux types de biomarqueurs. La plupart des biomarqueurs cliniques ont en effet été validés par diverses études parallèles faisant intervenir un nombre important de patients et des jeux de données différents. La phase de sélection est donc globalement terminée.

A l'opposé, elle est encore pleinement d'actualité pour les gènes. Cette fois, les études ne font intervenir qu'un nombre limité de patients, et n'ont pas encore été validées sur d'autres jeux de données. Les résultats obtenus sur une seule étude sont souvent considérés comme une vérité absolue, et ceci sans validation externe.

Comme exposé dans la partie 2.3.4, la question de la validation des études transcriptomiques a fait l'objet de travaux récents, montrant l'impact du processus de sélection au niveau du FDR et de la reproductibilité des études. La même question de la validation se pose pour les modèles de survie, qui sont soumis à un double enjeu : la sélection des "bons" gènes, et une estimation correcte de leur effet sur la survie.

L'objectif de ce travail a été de quantifier l'optimisme relatif aux variables transcriptomiques d'une part, et aux variables clinico-biologiques classiques d'autre part, quand les deux types de variables sont introduits dans un même modèle de survie. Il a fait l'objet d'une communication orale aux 27^{eme} Conférences de l'ISCB (International Society for Clinical Biostatistics) et l'article correspondant a été soumis [111].

5.2 Matériel et méthodes

Pour répondre à la question posée, le travail s'est décomposé en plusieurs étapes, qui sont illustrées dans l'organigramme de la figure 5.1.

Ces étapes seront reprises et développées une à une.

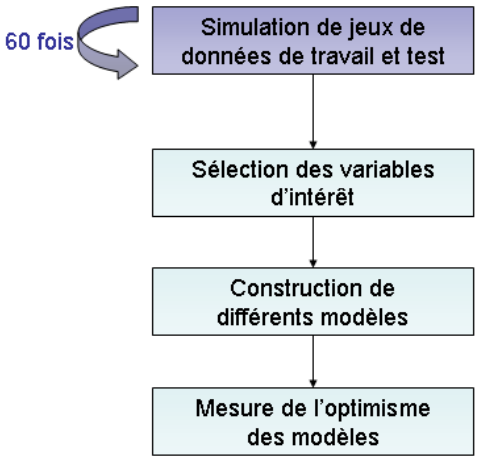


FIG. 5.1: *Etapes de l'analyse*

5.2.1 Simulation des jeux de données

Les jeux de données ont été simulés comme décrit dans la partie 3.2.2. Trois paramètres interviennent : le nombre n de patients inclus dans l'étude, le nombre de gènes p_1 reliés à la survie, et le nombre total de gènes étudiés, p.

5.2.2 Sélection des variables d'intérêt

Nous avons choisi d'employer le modèle de Cox pour analyser les données de survie ainsi simulées. Un bref rappel de ce modèle et des notations utilisées est présenté ci-dessous.

5.2.2.1 Le modèle de Cox

Soit X une matrice (n, p) exprimant p variables (ici les gènes) pour n individus, telle que $X = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, où $\mathbf{x}_i$ est le vecteur contenant les niveaux d'expression des p gènes de l'individu i. t_i, $i = 1, .., n$ sont les temps de suivi pour chacun des individus. d_i, $i = 1...n$ sont les indicateurs de l'état observé du patient au temps t_i, avec $d_i = 1$ si le décès a déjà été observé au temps t_i, et $d_i = 0$ si l'individu est toujours vivant au temps t_i (donnée censurée). Le modèle de Cox à taux proportionnel relie le risque instantané de décès $\lambda(t, X)$ et la matrice de covariables X d'un individu par le modèle suivant :

$$\lambda(t, X) = \lambda_0(t) exp(\boldsymbol{\beta}' X)$$

où $\lambda_0(t)$ est le risque de base, $\boldsymbol{\beta} = \{\beta_1, ..., \beta_p\}$ est le vecteur de paramètres associé aux p variables. Le vecteur $\boldsymbol{\beta}$ optimal est celui qui maximise la vraisemblance du modèle ci-dessus, appelée vraisemblance partielle de Cox (PL) et définie par :

$$PL(\boldsymbol{\beta}) = \prod_{k \in D} \frac{exp(\boldsymbol{\beta'} \boldsymbol{x}_k)}{\sum_{j \in R_k} exp(\boldsymbol{\beta'} \mathbf{x}_j)}$$

où D est l'ensemble des indices de l'événement et R_k l'ensemble des individus à risque au temps t_k. On pose $l(\beta) = -logPL(\beta)$, la log-vraisemblance partielle du modèle.

Dans le cas des gènes, le nombre de variables est trop important pour obtenir la forme analytique de la vraisemblance. Une solution à ce problème est d'utiliser des méthodes de régularisation, comme évoqué dans la partie 2.3.3. Parmi les méthodes proposées dans la littérature, nous avons retenu la méthode du "gradient seuillé", dite TGD pour "Threshold Gradient Descent path". Initialement proposée par Friedman et Popescu [74], cette méthode a été adaptée au modèle de Cox dans le cadre des biopuces par Gui et Li [75].

5.2.2.2 Méthode du gradient

Le principe de la méthode du gradient est de construire le vecteur des paramètres $\boldsymbol{\beta}$ de manière itérative. Les valeurs des paramètres à chacune des itérations définissent un chemin dans l'espace des paramètres. Ce chemin est défini par un point initial, un point final, et une formule définissant le passage d'un point du chemin au suivant. Le point initial correspond au modèle nul, tous les paramètres étant à zéro ; le point final correspond au modèle incluant toutes les variables. Le déplacement se fait dans la la direction opposée au gradient d'une fonction de perte choisie, l'erreur quadratique moyenne par exemple pour le modèle de régression linéaire. Chaque pas est défini de la manière suivante :

$$\hat{\boldsymbol{\beta}}(\nu + \Delta\nu) = \hat{\boldsymbol{\beta}}(\nu) + \Delta\nu.\mathbf{g}(\nu)$$

Le paramètre ν, initialement nul, contrôle le nombre de pas ; $\Delta(\nu) > 0$ contrôle la largeur de ce pas ; enfin, $\mathbf{g}(\nu)$ est le gradient négatif de la fonction de perte à l'étape ν.

Avec cette méthode, les paramètres estimés pour des variables corrélées sont peu dispersés. D'autres méthodes de régularisation, telles la régression ridge basée sur la pénalité de type $L2$,

conduisent à ce même phénomène. A l'inverse, les méthodes basées sur la pénalité de type $L1$ produisent des estimations de paramètres très dispersées.

Dans l'objectif de construire des modèles intermédiaires entre ces deux extrêmes, et permettre une hétérogénéité des paramètres estimés, Friedman et Popescu [74] proposent une méthode de "gradient généralisé". La construction du vecteur de paramètres est cette fois définie de la manière suivante :

$$\hat{\boldsymbol{\beta}}(\nu + \Delta\nu) = \hat{\boldsymbol{\beta}}(\nu) + \Delta\nu.\mathbf{h}(\nu)$$

$\mathbf{h}(\nu)$ définit la direction du chemin dans l'espace des paramètres, tangente à $\hat{\boldsymbol{\beta}}(\nu)$, et est définie par $\mathbf{h}(\nu) = \{f_j(\nu).g_j(\nu)\}_1^p$.
Les termes $\{f_j(\nu)\}_0^p \leq 0$ permettent de pondérer les composantes du gradient. L'introduction de cette pondération permet la diversification des paramètres estimés. Sa définition conduit à différentes méthodes, dont celle du "gradient seuillé" TGD ("Threshold Gradient Descent"). Dans ce cas :

$$f_j(\nu) = I[|g_j(\nu)| \geq \tau.\max_{1\leq k\leq p}|g_k(\nu)|]$$

$I[.]$ correspond au symbole de Kronecker, et $\tau \in [0, 1]$ contrôle la diversité des paramètres estimés. A travers $\mathbf{f}(\nu)$, les coefficients mis à jour à chaque étape dépendent de la valeur de τ.

La figure 5.2, issue du rapport technique de Friedman et Popescu [74], illustre le rôle du paramètre τ. Des données ont été simulées pour 150 observations telles que sur 10 000 variables, seules les 100 premières sont prédictives dans un modèle de régression linéaire. Les estimations des paramètres en bleu sont celles des variables simulées comme prédictives, tandis que celles en rouge sont celles des variables correspondant à du bruit. Pour $\tau = 0$, toutes les variables sont retenues dans le modèle final et les estimations sont peu dispersées. Toutes les variables prédictives sont conservées dans le modèle, mais au prix de la conservation de variables de bruit. Avec $\tau = 1$, un nombre restreint de variables est cette fois retenu. L'essentiel des variables retenues sont effectivement prédictives, mais toutes les variables prédictives n'ont pas été retenues. Un seuil $\tau = 0.6$ permet un compromis entre ces extrêmes : davantage de variables prédictives sont retenues, tout en conservant un nombre de variables de bruit faible.

Le seuil permet ainsi de définir un compromis entre le nombre de vrais positifs et de faux positifs conservés dans le modèle final. On peut également noter que plus il y a de variables sélectionnés, plus l'estimation de leur coefficient est faible. Ce paramètre influence donc égale-

ment la force du rétrécissement des coefficients des variables introduites dans le modèle final.

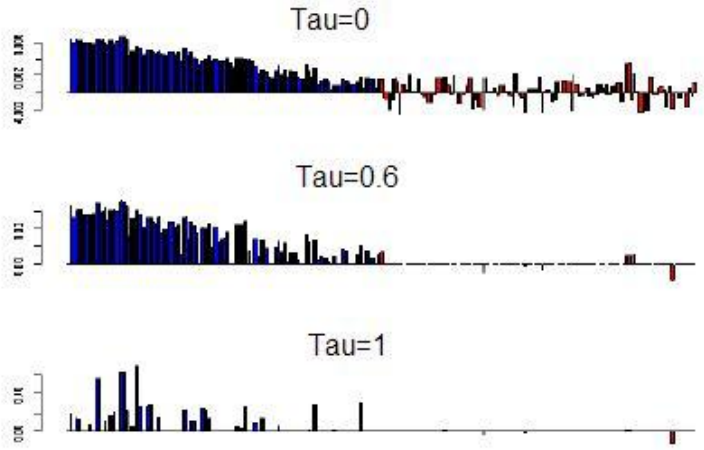


FIG. 5.2: *Estimation des 200 premiers paramètres du modèle de régression pour différentes valeurs de τ. Les estimations des paramètres en bleu sont celles des variables prédictives, tandis que celles en rouge sont celles des variables correspondant à du bruit.*

5.2.2.3 Adaptation au modèle de Cox

Dans le contexte des biopuces, Gui et Li [75] ont adapté le modèle du TGD au modèle de Cox. Les auteurs ont choisi comme fonction de perte la log-vraisemblance partielle $l(\beta) = -logPL(\beta)$ définie plus haut. Dans ce cas, le gradient est défini par $\partial l/\partial\beta$ et la construction du vecteur de paramètres se fait dans la direction opposée au gradient : $\mathbf{g} = -\partial l/\partial\beta$

L'algorithme de cette méthode est le suivant :

1. $\beta(0) = 0$ et $\nu = 0$
2. Calcul du gradient de la fonction de perte $\mathbf{g} = -\partial l/\partial\beta$ pour le β courant.
3. Calcul de de $f_j(\nu) = I[|g_j(\nu)| \geq \tau.\max_{1\leq k\leq p}|g_k(\nu)|]$
4. Mise à jour du vecteur de paramètres : $\hat{\boldsymbol{\beta}}(\nu + \Delta\nu) = \hat{\boldsymbol{\beta}}(\nu) + \Delta\nu.\mathbf{h}(\nu)$
5. Répétition des étapes 2 à 4 jusqu'à convergence de l'estimation de β.

La valeur de ν optimale est celle qui minimise la log-vraisemblance partielle cross-validée. Au terme du processus itératif, seuls les gènes estimés comme reliés à la survie ont un coefficient non nul, et seront donc ainsi sélectionnés.

Nous avons choisi $\tau = 0.8$, comme proposé dans l'application de Gui et Li [75] de la méthode au jeu de données de Rosenwald sur 240 patients atteints d'un démembrement des lymphomes

B diffus à grandes cellules (DLBCL). Étant donné le nombre important de gènes non informatifs attendus sur une biopuces, ce choix nous a semblé un bon compromis pour conserver dans le modèle le maximum de vrais positifs sans conserver trop de faux positifs.

5.2.2.4 Bilan

Appliquée aux gènes, la méthode du TGD combine ainsi la sélection des gènes d'intérêt et l'estimation de leur effet sur la survie. Nous avons utilisé cette méthode uniquement comme méthode de sélection, comme cela sera explicité par la suite (cf partie 5.4).
Pour les variables clinico-biologiques classiques en revanche, et pour tenir compte du fait qu'elles ne sont plus sujettes à sélection, nous les avons toutes deux introduites dans les modèles, que leur apport soit ou non significatif sur le jeu de données considéré.

5.2.3 Construction des modèles

Trois modèles ont été considérés. Les variables cliniques sont regroupées dans une matrice $(n, 2)$ notée X_C. Les gènes sélectionnés par le gradient sont regroupés dans une matrice (n, k) notée X_T, où k est le nombre de gènes sélectionnés sur un jeu de données particulier.

$$\lambda(t) = \lambda_{C0}(t) \exp(\boldsymbol{\alpha}_C X_C) \tag{5.1}$$

$$\lambda(t) = \lambda_{T0}(t) \exp(\boldsymbol{\alpha}_T X_T) \tag{5.2}$$

$$\lambda(t) = \lambda_0(t) \exp(\boldsymbol{\gamma}_T X_T + \boldsymbol{\gamma}_C X_C) \tag{5.3}$$

Le modèle défini par l'équation (5.1) sera nommé par la suite "modèle clinique". Il introduit les variables cliniques seules dans un modèle de Cox multivarié. Le modèle (5.2), qualifié de "modèle transcriptomique", introduit dans un modèle de Cox multivarié les gènes qui ont été sélectionnés par le gradient. Dans ce modèle, les coefficients sont réestimés et ne correspondent pas aux estimations obtenues par la méthode du gradient. Le modèle défini par l'équation (5.3) ajuste simultanément les deux types de variables.

5.2.4 Mesure de l'optimisme des modèles

5.2.4.1 Information apportée par un modèle

Afin de mesurer puis de comparer l'optimisme des modèles décrits ci-dessus, un outil de mesure de la qualité prédictive de ces modèles pronostics est nécessaire. Cet outil doit permettre d'évaluer la capacité du modèle à prédire le devenir de nouveaux patients qui n'ont pas contribué à sa construction. La quantité de l'information apportée par les variables introduites dans le modèle, qui correspond à la part de variance expliquée par ce modèle, joue un rôle important pour cela. La variabilité présente dans les données et celle expliquée par le modèle est en effet le reflet de la robustesse de ce dernier et de l'efficacité de son utilisation sur d'autres jeux de données. Partant de ce fait, nous avons choisi de nous baser sur l'information apportée dans les modèles par les deux types de variables pour comparer ensuite l'optimisme relatif à chacun des types de variables.

Une mesure classiquement utilisée pour quantifier la part de variabilité expliquée par un modèle de régression linéaire est le coefficient de détermination R^2. Soit le modèle de régression linéaire défini par :

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \epsilon_i$$

y_i est la variable de réponse pour l'individu i, $\{\beta_0, \ldots, \beta_p\}$ les coefficients à estimer, $\{X_1, \ldots, X_p\}$ sont les vecteurs décrivant les p variables explicatives, et ϵ_i est le terme d'erreur.

Le coefficient de détermination R^2 s'interprète comme la part de la variance de la variable Y expliquée par la régression ; il varie entre 0 et 1 et est défini comme suit :

$$R^2 = SCE/SCT = 1 - SCR/SCT$$

Dans cette formule, $SCT = \sum_i (y_i - \bar{y})^2$ reflète la variance totale ; $SCE = \sum_i (y_i - \hat{y}_i)^2$ reflète la part de variance expliquée par le modèle ; $SCR = \sum_i (\hat{y}_i - \bar{y})^2$ reflète la part de variance résiduelle non expliquée par le modèle.

En survie, les observations sont des temps tandis que le modèle prédit un risque instantané de décès ; la formulation proposée pour le R^2 en régression linéaire ne peut être utilisée directement et doit être adaptée.

5.2.4.2 ρ^2_{IG} de Kent et O'Quigley

Kent et O'Quigley ont proposé une reécriture du R^2 basée sur la théorie de l'information [112, 113], et sur la notion d'entropie.

L'entropie d'une variable aléatoire permet de mesurer la quantité d'incertitude reliée à cette variable. L'entropie est ainsi une mesure de l'information : plus il y a d'incertitude sur cette variable, plus elle apporte d'information.

Soit alors un modèle paramétrique $f(t|z)$ de paramètre θ. T est la variable à expliquer et Z correspond aux variables explicatives. En considérant qu'une distribution de probabilité représente la connaissance sur une variable aléatoire, on peut mesurer l'information apportée par le paramètre θ par $I(\theta) = E\{\log f(T|Z;\theta\})$, cette formulation reposant sur la définition de l'entropie.

En utilisant la distribution empirique de (T, Z), cette information peut être estimée par [112] :

$$\hat{I}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathcal{T}}\{\log f(t|Z_i;\theta\}f(t|Z_i;\hat{\beta})dt$$

où n est la taille de l'échantillon considéré, $\mathcal{T}$ est le domaine de définition de T, et $\hat{\beta}$ est une estimation du maximum de vraisemblance des paramètres du modèle.

La variabilité totale de T, notée $D(T)$, est mesurée par l'entropie du modèle sans covariables, c'est à dire le modèle nul où $\theta = 0$. L'entropie étant insensisble à toute transformation monotone, on peut la réécrire comme $D(T) = exp\{-2(I(\theta = 0))\}$.

La variabilité résiduelle, notée $D(T|Z)$, quant à elle est mesurée par l'entropie conditionnelle de T sachant Z. On peut l'écrire après transformation comme $D(T|Z) = exp\{-2(I(\theta = \beta))\}$.

La proportion de variation expliquée par le modèle peut alors s'écrire comme :

$$\begin{aligned}
\rho^2_{IG} &= \frac{D(T) - D(T|Z)}{D(T)} \\
&= 1 - \frac{D(T|Z)}{D(T)} \\
&= 1 - \frac{exp\{-2(I(\beta))\}}{exp\{-2(I(0))\}} \\
&= 1 - exp\{-2(I(\beta) - 2I(0))\}
\end{aligned}$$

En théorie de l'information, la quantité $\Gamma(\beta) = 2\{I(\beta) - I(0)\}$ est appelée gain d'information, ou entropie relative. Elle quantifie la différence entre l'information correspondant à $\theta = \beta$ et celle correspondant à $\theta = 0$. On parle également de distance de Kullback-Leibler.

On peut noter les propriétés suivantes : $0 \leq \rho^2_{IG} < 1$ avec $\rho^2_{IG} = 0$ pour le modèle nul et $\rho^2_{IG} \to 1$ quand tous les paramètres tendent vers l'infini.

Cette réécriture du R^2 s'applique à toute fonction de distribution paramétrique. Dans le cas où f correspond à la fonction de densité de la loi normale, et où les paramètres sont estimés par les moindres carrés ou par le maximum de vraisemblance, $\hat{\rho}^2_{IG}$ coincide avec le coefficient de détermination R^2.

Les paramètres du modèle de Cox sont invariants par une transformation monotone croissante du temps. En utilisant une transformation adaptée du temps T, la distribution conditionnelle de T peut être décrite par une fonction de Weibull. En réécrivant ainsi le modèle de Cox sous une forme paramétrique, le gain d'information et donc ρ^2_{IG} peuvent être estimés aisément.

Maucort-Boulch *et al.* ont étudié plus particulièrement le comportement de cet outil face à la censure. Puisque son calcul ne nécessite pas l'utilisation de la vraisemblance partielle de Cox, il ne fait pas intervenir la censure, ce qui en fait un outil de choix pour les modèles survie. Nous avons donc choisi le ρ^2_{IG} proposé par Kent et O'Quigley comme outil de comparaison de l'optimisme relatif aux modèles décrits ci-dessus. Par la suite, nous noterons plus simplement $\rho^2 = \rho^2_{IG}$.

5.2.4.3 Application

Ce sont les valeurs de ρ^2 pour les différents modèles qui ont permis de comparer l'optimisme relatif aux variables cliniques et transcriptomiques.

Nous parlerons de ρ^2 ajusté ou non selon qu'il a été calculé respectivement sur les modèles incluant les deux types de variables ou uniquement l'un d'entre eux. Attention, le sens donné ici au ρ^2 ajusté ne doit pas être confondue avec la définition classique utilisé en régression, où le R^2 ajusté correspond à une réécriture du R^2 qui tient compte du nombre de paramètres explicatives introduites dans le modèle de régression.

Les valeurs de ρ^2 ont été calculées à trois niveaux pour chacun des modèles, comme décrit ci-dessous.

A partir des coefficients simulés, ρ^2_{exp}. Les coefficients théoriques fixés dans les simulations sont directement utilisés pour le calcul du ρ^2. C'est la valeur de ρ^2 qu'on s'attendrait à observer si le modèle estimé correspondait à celui simulé (l'indice "exp" est employé pour "expected"). Les données ayant été générées à partir d'un modèle ajusté, ρ^2_{exp} est un ρ^2 ajusté.

Sur le jeu de travail, ρ^2_{train}. En ce qui concerne les gènes, ils sont sélectionnés sur le jeu de travail et leurs coefficients sont estimés sur ce même jeu. Les coefficients des modèles 5.1 et 5.3 sont également estimés sur le jeu de travail.

Sur le jeu de test, ρ^2_{test}. Les coefficients des gènes qui ont été sélectionnés sur le jeu de travail sont réestimés sur le jeu test pour les modèles 5.2 et 5.3. Par la suite nous utiliserons $\bar{\rho}^2_{test}$, qui correspond à la moyenne des ρ^2 estimés sur les 50 jeux tests.

Il est nécessaire de revenir plus en détail sur le calcul de ρ^2_{test}, car il n'est pas classique et son calcul a été adapté à notre problématique et à l'outil de mesure. La démarche habituelle pour mesurer l'optimisme est la suivante : 1- Les coefficients du modèle de prédiction sont estimés sur un jeu de travail ; 2- Le modèle est évalué sur un jeu test en utilisant les coefficients estimés sur le jeu de travail ; 3- Les qualités prédictives obtenues sur chacun des jeux sont comparées. En faisant cela, on mesure l'optimisme relié à l'estimation des paramètres et à la trop forte adéquation du modèle sur le jeu de travail. Cependant, ce processus ne tient pas compte de l'étape de sélection des variables.

Dans le modèle de Cox, l'information sur la censure est contenue dans l'estimation $\hat{\beta}$ des coefficients du modèle. Le calcul du ρ^2 ne dépend que de la valeur des $\hat{\beta}$ et celle des variables associées. En utilisant sur un autre jeu de données les estimations obtenues sur le jeu de travail, on ferait l'hypothèse que les variables sélectionnées sur le jeu de travail sont également prédictives sur le jeu test. En effet, le ρ^2 permet de quantifier l'information apportée par les variables introduites dans le modèle. Cette information est contenue dans les paramètres β. En utilisant ces paramètres sur un autre jeu de données, ici le jeu de données test, on attribue aux variables de ce jeu de données l'information qui avait été obtenue par ces mêmes variables dans le jeu de travail. Dans ce cas, c'est l'optimisme dû à l'estimation des paramètres qui est mesuré, et celui dû à la sélection des variables correspondantes est omis.

Pour les variables cliniques, considérées comme validées, l'optimisme dû à la sélection n'intervient pas. Dans ce cas, les $\rho_t^2 est$ moyens obtenus en réestimant ou non les coefficients des deux variables cliniques sont similaires.

Les gènes, en revanche, sont en pleine phase de sélection. La mesure de l'optimisme doit évaluer simultanément l'effet de la sélection des gènes et de l'estimation de leur effet prédictif sur de nouveaux jeux de données. Pour cette raison, les coefficients des gènes sélectionnés sur le jeu de travail ont été réestimés sur le jeu test. Ce sont ces nouveaux coefficients qui ont été utilisés pour le calcul de ρ_{test}^2. Plus particulièrement, nous avons travaillé sur $\bar{\rho}_{test}^2$, qui correspond à la moyenne des ρ^2 estimés sur l'ensemble des 50 jeux tests générés pour un même jeu de travail.

Le tableau 5.1 résume ces notions :

Paramètres utilisés	Signification
$X_{train}\beta_{train}$	Quantité d'information contenue par les gènes sélectionnés sur le jeu où ils ont été sélectionnés
$X_{test}\beta_{train}$	Quantité d'information qui serait contenue par les gènes sélectionnés si les β_{train} étaient les bons
$X_{test}\beta_{test}$	Quantité d'information effectivement contenue dans le jeu test par les gènes sélectionnés sur le jeu de travail

Tab. 5.1: *Signification des ρ^2 selon les paramètres utilisés pour leur calcul.*

Ce sont ensuite les différences entre les ρ^2 qui nous ont permis de quantifier l'optimisme :

$$\Delta_{TrExp} = \frac{\sum_{i=1}^{60} \left(\rho_{train,i}^2 - \rho_{exp,i}^2\right)}{60} \tag{5.4}$$

$$\Delta_{TeExp} = \frac{\sum_{i=1}^{60} \left(\bar{\rho}_{test,i}^2 - \rho_{exp,i}^2\right)}{60} \tag{5.5}$$

$$\Delta_{TrTe} = \frac{\sum_{i=1}^{60} \left(\rho_{train,i}^2 - \bar{\rho}_{test,i}^2\right)}{60} \tag{5.6}$$

Comparaison de ρ_{train}^2 et ρ_{exp}^2, Δ_{TrExp} (Équation 5.4) : Cette variable exprime la différence entre la quantité d'information prédictive réelle d'un jeu de données et celle observée sur le jeu avec lequel les variables ont été sélectionnées. On peut ainsi évaluer l'optimisme introduit par l'estimation de l'effet sur la survie des gènes sélectionnés.

Comparaison de $\bar{\rho}^2_{test}$ et ρ^2_{exp}, Δ_{TeExp} (Équation 5.5) : Cette variable exprime la différence entre la quantité d'information prédictive réelle d'un jeu de données et celle mesurée sur le jeu test avec les variables sélectionnées sur le jeu de travail. On peut cette fois évaluer l'optimisme introduit par l'étape de sélection des variables.

Comparaison de ρ^2_{train} et $\bar{\rho}^2_{test}$, Δ_{TrTe} (Équation 5.6) : Cette variable exprime la différence entre la quantité d'information prédictive détectée sur un jeu de données, et celle obtenue avec les mêmes variables sur un autre jeu de données. On a ainsi une mesure de l'erreur qui est faite en considérant que les gènes sélectionnés sur un jeu de données sont les bons et auraient donc le même pouvoir prédictif sur d'autres jeux de données. On peut cette fois évaluer l'optimisme introduit simultanément par la sélection des variables et par l'estimation de leurs coefficients.

5.3 Résultats

Grâce à ces outils, l'influence sur l'optimisme de trois paramètres a été évaluée : le nombre d'individus n, et de gènes p, introduits dans l'étude, ainsi que le nombre p_1 de gènes réellement reliés à la survie. Pour une meilleure lisibilité, les résultats sont présentés sous la forme de graphiques. Sur ces graphiques figurent également les intervalles de confiance empiriques calculés sur l'ensemble des 60 jeux de données simulés.

5.3.1 Influence du nombre de patients

Les résultats ont été obtenus avec un nombre total de gènes $p = 1000$ fixé, pour un nombre variable de gènes sous H_1, $p_1 = \{5, 10, 20\}$.

Comparaison de ρ^2_{train} et $\bar{\rho}^2_{test}$, Δ_{TrTe} : La figure 5.3 montre Δ_{TrTe} (différence entre ρ^2_{train} et ρ^2_{test}) en fonction du nombre de patients, pour les modèles transcriptomiques et cliniques. Pour le modèle transcriptomique, Δ_{TrTe} diminue quand n augmente, sans jamais atteindre zéro, et ceci quelle que soit la valeur de p_1. La largeur des intervalles de confiance est globalement constante quel que soit le nombre de patients. La taille de l'intervalle de confiance pour 400 patients et $p1 = 20$ s'explique par le fait qu'avec ce jeu de paramètres, la méthode du gradient converge pour moins de jeux de données.

En ce qui concerne le modèle clinique, Δ_{TrTe} est très proche de zéro, et ceci d'autant plus que le nombre de patients augmente. La largeur des intervalles de confiance diminue avec n et est nettement plus faible que pour le modèle transcriptomique.

Ces résultats indiquent que le pouvoir prédictif des gènes sélectionnés sur un jeu de données est surestimé par rapport au pouvoir prédictif qu'ils auraient sur un autre jeu de données. Pour les variables cliniques en revanche, cette surestimation est quasiment nulle. Selon la nature des variables, les deux modèles ont un comportement très différent ; ils ne peuvent donc pas être interprétés de la même manière. Les mêmes résultats ont été obtenus pour le modèle ajusté.

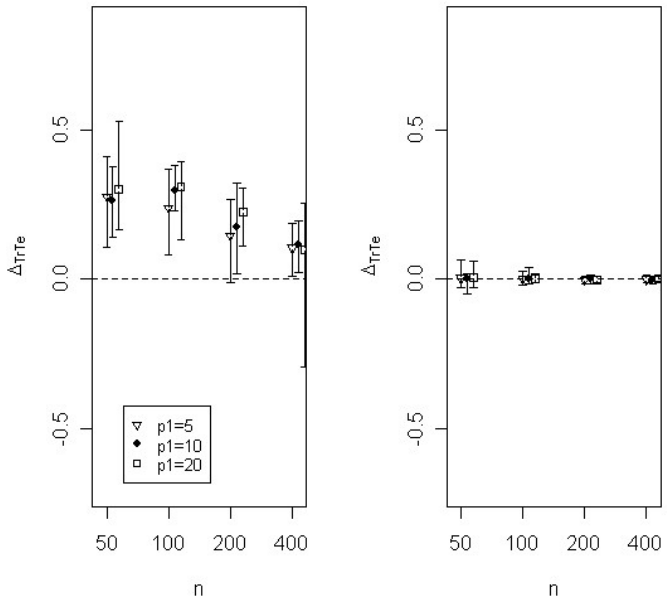


FIG. 5.3: *Évolution de Δ_{TrTe} en fonction du nombre de patients n pour les modèles transcriptomique (à gauche), et clinique (à droite) - $p = 1000$ gènes. Les intervalles de confiance sont représentés par les segments.*

Comparaison de ρ^2_{train} et ρ^2_{exp}, Δ_{TrExp} : La figure 5.4 montre Δ_{TrExp} pour les deux types de variables dans le modèle ajusté. Pour les variables transcriptomiques, Δ_{TrExp} tend vers zéro quand le nombre de patients augmente, le pouvoir prédictif observé sur le jeu de travail tendant alors vers celui qui est attendu. Ces résultats indiquent que si les jeux de données sont trop petits, le pouvoir prédictif qu'on croit être contenu dans les gènes sélectionnés est en réalité surestimé. La méthode du gradient conduit à sélectionner un nombre de gènes supérieur à p_1 ; ces faux positifs font gonfler ρ^2_{train}, donnant l'illusion que le pouvoir prédictif des gènes sélectionnés augmente alors que cette augmentation est due à du bruit.

Cette fois, le nombre de gènes sous H_1 a un effet : Δ_{TrExp} est d'autant plus grand que p_1 est petit. Par nature, ρ^2 dépend du nombre de variables introduites dans le modèle. De ce fait, ρ^2_{exp} augmente avec p_1. Par contre, le nombre de gènes sélectionnés ne dépend que très peu de p_1, et reste de l'ordre de la dizaine. Δ_{TrExp} est donc d'autant plus grand que p_1 est petit puisque dans ce cas, le modèle obtenu sur le jeu de travail contient plus de variables que le modèle théorique.

Les résultats pour les variables cliniques sont très différents ; le pouvoir prédictif des modèles est cette fois sous-estimé, ceci d'autant plus que p_1 est grand. Cette observation s'explique par le biais dû aux covariables manquantes quand les modèles sont mal spécifiés [114]. En effet, l'omission de variables explicatives dans un modèle conduit à sous-estimer l'effet des variables non omises. Le nombre de vrais positifs étant faible quel que soit le nombre de gènes p_1 reliés à la survie, il manque d'autant plus de gènes dans le modèle que p_1 est grand. Cette propriété a également des effets sur le comportement du modèle transcriptomique. La méthode du gradient manque certains gènes d'intérêt. Puisque le nombre de gènes sélectionnés varie peu, elle en manque d'autant plus que p_1 augmente. Par contre, l'introduction de faux positifs n'introduit pas de biais dans l'estimation de l'effet des vrais positifs.

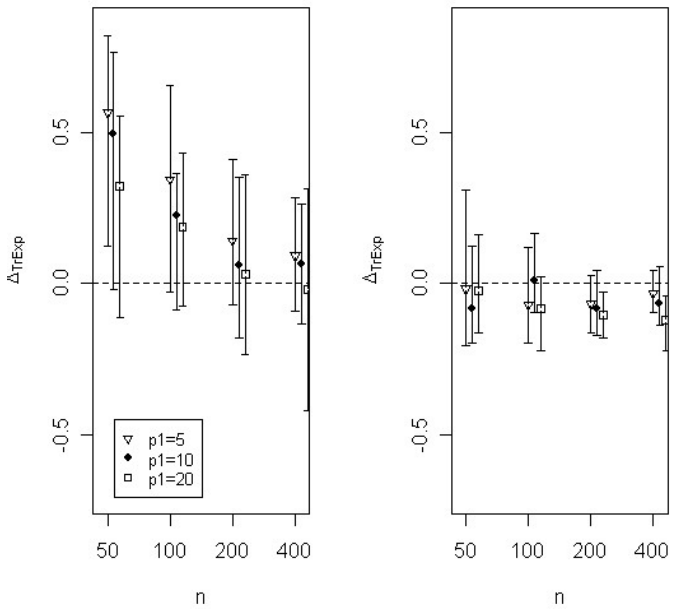


FIG. 5.4: *Évolution de Δ_{TrExp} en fonction du nombre de patients n pour les variables transcriptomiques (à gauche), et cliniques (à droite) des modèles ajustés- $p = 1000$ gènes. Les intervalles de confiance sont représentés par les segments.*

Comparaison de $\bar{\rho}^2_{test}$ et ρ^2_{exp}, Δ_{TeExp} : Les résultats de la figure 5.5 montrent que sur les jeux tests, les gènes sélectionnés sur le jeu de travail sont incapables de restituer l'information

prédictive théorique du jeu de données. En effet, hormis pour 50 patients, les valeurs de Δ_{TeExp} sont négatives, et ceci d'autant plus que le nombre de gènes p_1 reliés à la survie est grand. En effet, la valeur de ρ^2 dépend du nombre de covariables sélectionnées et introduites dans le modèle, et le nombre de gènes sélectionnés ne dépendant pas de p_1, il manque d'autant plus de variables que p_1 est petit. Avec 200 et 400 patients, le nombre de vrais positifs augmente, ce qui explique la légère augmentation de Δ_{TeExp}.

Comme pour les variables issues du transcriptomique, les résultats obtenus pour les variables cliniques sont atypiques pour 50 patients. Pour un nombre de patients plus important, la sous-estimation de l'effet des variables cliniques s'explique à nouveau par le biais déjà évoqué ci-dessus.

Les valeurs positives observées pour 50 patients signifient que les coefficients estimés sur les jeux tests sont plus élevés que ceux attendus, alors même que l'on s'attend à ce que les gènes sélectionnés soient des faux positifs. Ceci peut s'expliquer par la forte variance des estimateurs des paramètres du modèle due à la petite taille de l'échantillon sur lequel ils sont estimés.

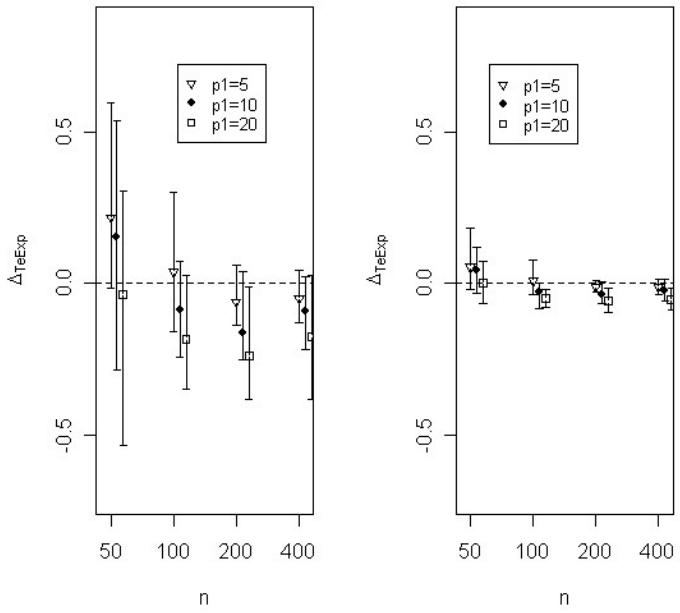


FIG. 5.5: *Évolution de Δ_{TeExp} en fonction du nombre de patients n pour les variables transcriptomique (à gauche), et cliniques (à droite) dans le modèle ajusté- $p = 1000$ gènes. Les intervalles de confiance sont représentés par les segments.*

Rôle des vrais positifs Nous avons également voulu étudier l'optimisme relatif aux vrais positifs (VP). Pour cela, nous avons comparé l'évolution des ρ^2 calculés sur les VP d'une part, et à tous les gènes sélectionnés d'autre part. Les résultats obtenus pour $p = 1000$ et $p1 = 10$ gènes

sont présentés sur la figure 5.6. Sur les jeux de travail, quand n augmente, ρ^2_{train} calculé à partir de l'ensemble des gènes sélectionnés varie peu (hormis pour $n = 50$), tandis que ρ^2_{train} calculé uniquement sur les VP augmente. Les ρ^2_{train} observés sur les jeux de travail correspondent donc à du bruit. En utilisant un seul jeu de données pour une étude, cela ne peut pas être mis en évidence, mais peut conduire à une mauvaise interprétation des résultats. Sur les jeux test, $\bar{\rho}^2_{test}$ est du même ordre de grandeur selon qu'il est calculé à partir de l'ensemble des gènes ou uniquement des VP, puisque les faux positifs n'apportent pas d'information. Avec trop peu de patients (50 ou 100), la méthode ne détecte aucun VP, d'où l'absence d'intervalles de confiance pour ces valeurs. Dans ce dernier cas, les valeurs de ρ^2_{train} et ρ^2_{test} semblent anormalement hautes, particulièrement pour 50 patients. Comme énoncé précédemment, ceci s'explique par la forte variance des estimateurs due à des échantillons de taille modeste.

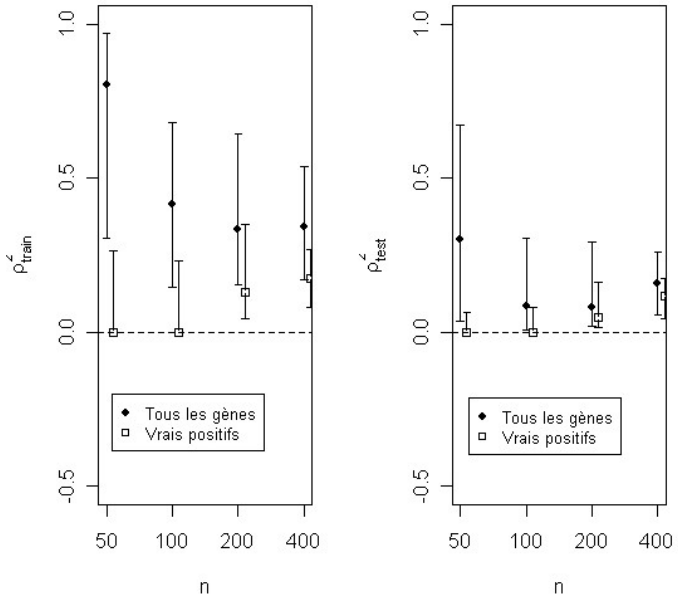


FIG. 5.6: *Évolution de ρ^2_{train} (à droite) et $\bar{\rho}^2_{test}$ (à gauche) en fonction de n calculés sur l'ensemble des gènes sélectionnés ou uniquement les vrais positifs - $p = 1000$ et $p_1 = 10$ gènes. Les intervalles de confiance sont représentés par les segments.*

5.3.2 Influence du nombre total de gènes

Les résultats ont été obtenus avec un nombre de patients $p = 100$ fixé, pour un nombre variable de gènes sous H_1, $p_1 = \{5, 10, 20\}$. Seuls les résultats concernant la comparaison entre ρ^2_{train} et $\bar{\rho}^2_{test}$ (Δ_{TrTe}) sont présentés, les autres n'apportant pas d'information complémentaire. En effet, le nombre de gènes sélectionnés par le gradient augmente avec le nombre total de gènes introduit dans l'étude. Par conséquent, quand le nombre total de gènes augmente, le nombre

de variables qui contribuent au calcul de ρ^2, et donc la valeur de ρ^2 augmente. Or, le nombre de variables qui contribuent au calcul de ρ^2_{exp}, lui, est constant à p_1 constant. De ce fait, les différences observées Δ_{TrExp} et Δ_{TeExp} en fonction de p sont peu interprétables, la part due respectivement aux VP et aux FP n'étant pas connue. Ceci n'était pas le cas pour les différences en fonction de n, le nombre de gènes sélectionnés par le gradient étant moins influencé par le nombre de patients inclus dans l'étude.

La figure 5.7 montre les résultats obtenus pour Δ_{TrTe} dans le cas du modèle transcriptomique. Δ_{TrTe} augmente avec le nombre total de gènes évalués par le modèle du gradient. Avec une trop faible proportion de gènes sous H_1, la méthode est incapable de les détecter. De ce fait, les gènes sélectionnés sur le jeu de travail sont des FP qui n'ont aucune valeur prédictive sur les jeux tests. On ne peut pas considérer que la valeur de p_1 joue un rôle étant donné le recouvrement des intervalles de confiance.

Concernant les variables cliniques, leur effet est sous-estimé. Ceci s'explique par le biais évoqué déjà ci-dessus, introduit par le fait qu'il manque dans le modèle des covariables liées à la survie.

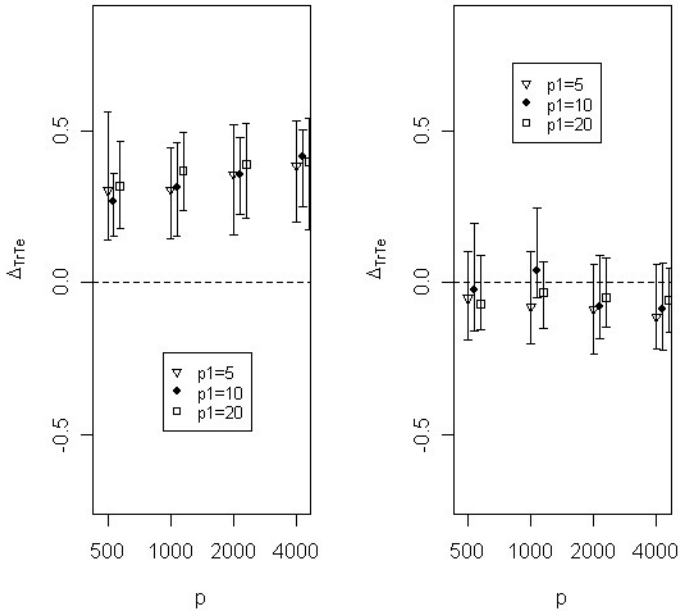


FIG. 5.7: *Évolution de Δ_{TrTe} en fonction du nombre de gènes p pour les modèles transcriptomique (à gauche), et clinique (à droite) - $n = 100$ patients. Les intervalles de confiance sont représentés par les segments.*

5.4 Remarques sur le gradient

La méthode du gradient allie les étapes de sélection des variables d'intérêt et d'estimation de leur effet sur la survie. L'utilisation de cette méthode nous a conduit à quelques remarques à son sujet.

5.4.1 Estimation des coefficients

Dans leur article, Gui et Li [75] appliquent la méthode du gradient à un jeu de données sur le lymphome à grandes cellules B[1] [47]. Avec un seuil de $\Gamma = 1$, quatre gènes sont sélectionnés, avec des coefficients qui ne dépassent pas 0.1, le plus faible des coefficients étant quasiment nul. Les auteurs suggèrent que les estimations sont d'autant plus faibles qu'il y a d'hétérogénéité dans les données.

Nous avons également constaté dans nos simulations que la méthode a tendance à trop corriger les paramètres, qui sont très sous-estimés par rapport aux valeurs théoriques simulées. Pour certains jeux de données les coefficients sont quasiment nuls (de l'ordre de 10^{-15}). Ceci entraîne deux cas de figure selon que l'estimation des coefficients dans un nouveau modèle de Cox multivarié des gènes ainsi sélectionnés reste faible ou non. Dans le premier cas, cela signifie que les gènes sont des faux positifs, tandis que dans le second ce sont des vrais positifs. Au niveau de l'estimation des coefficients, la méthode du gradient ne permet pas de distinguer les vrais des faux positifs. Nous avons étudié l'évolution du pourcentage de jeux de données sur lesquels les gènes sélectionnés par le gradient ont des coefficients "raisonnablement non nuls" (c'est à dire pas de l'ordre de 10^{-15}) lorsqu'ils sont réestimés sur le jeu de travail qui a permis leur sélection. Nous les avons qualifié de "jeux informatifs". La figure 5.8 représente cette évolution en fonction du nombre de patients inclus dans l'étude. Chaque point correspond à la médiane du nombre de jeux informatifs obtenus pour les différentes valeurs possible de gènes sous $H1$ ($p_1 = \{5, 10, 20\}$).

On observe qu'en augmentant la taille de l'étude, l'estimation de l'effet des VP s'améliore. Ce graphique est une autre manière d'appréhender l'évolution des VP.

Dans un premier temps, le modèle issu de la méthode du gradient avait été considéré en conservant les estimations des coefficients. L'estimation des paramètres étant trop faible, ce

[1] Ce jeu de données contient les niveaux d'expression de 7399 gènes, pour 240 patients répartis en un jeu de travail (160 patients) et un jeu test (80 patients).

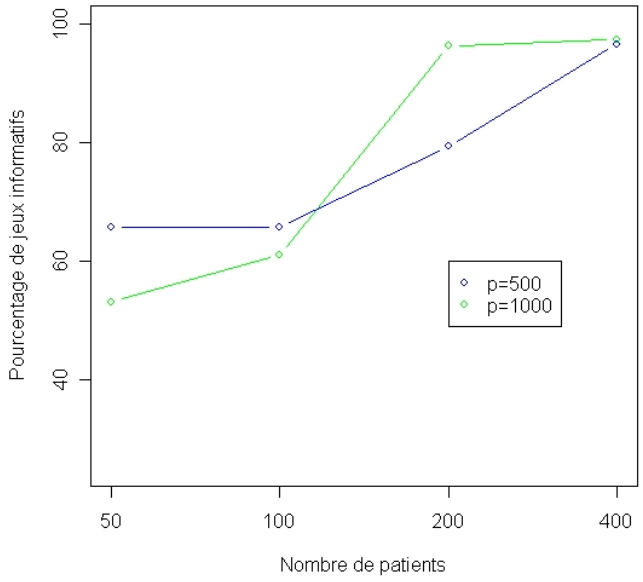


FIG. 5.8: *Évolution du nombre de jeux informatifs en fonction du nombre de patients inclus dans l'étude pour 500 (en bleu) et 1000 (en vert) gènes.*

modèle n'a pas été retenu pour la suite du travail, le gradient ayant uniquement été utilisé à fin de sélection.

5.4.2 Sélection des variables d'intérêt

Le nombre de gènes sélectionnés dépend du nombre de patients et de gènes inclus dans l'analyse, comme illustré sur la figure 5.9. Elle montre la médiane sur l'ensemble des soixante jeux de travail simulés, du nombre de gènes sélectionnés en fonction du nombre de patients inclus dans l'étude. Les intervalles de confiance des médianes sont également indiqués. En raison de l'étendue des intervalles de confiance, les deux graphiques ne sont pas à la même échelle. Deux modèles ont été considérés : le modèle 1 (en noir) correspond au gradient "classique" décrit précédemment (cf partie 5.2.2), tandis que le modèle 2 (en rouge) correspond à une modification de la méthode du gradient que nous avons proposée. Elle contraint l'introduction des variables clinico-biologiques classiques dans le modèle. Pour reproduire le fait que les variables cliniques ne sont plus en phase de sélection, leurs coefficients sont mis à jour à chaque itération, indépendamment de la mise à jour des coefficients des gènes.

En vert et en bleu figurent le nombre de vrais positifs sélectionnés respectivement par les modèles 1 et 2. La ligne horizontale en pointillés indique le nombre de gènes simulés sous $H1$.

Une première constatation est que le nombre de gènes sélectionnés est très variable d'un jeu de données à l'autre, comme l'indique la largeur des intervalles de confiance. Le modèle 2 conduit à identifier davantage de vrais positifs, mais au prix d'un plus grand nombre de gènes sélectionnés. Les médianes du complément du FDR (1-FDR) pour les mêmes soixante jeux de données sont représentées en fonction du nombre de patients sur la figure 5.10. Cette proportion correspond à la proportion de VP détectés parmi les gènes sélectionnés. La ligne horizontale en pointillés correspond à (1-FDR) de 50%. Il apparaît tout de suite que la proportion de gènes sous $H1$ détectés parmi les gènes sélectionnés est faible, puisqu'elle ne dépasse pas 50%, ce qui correspond à un FDR élevé. La méthode est d'autant plus performante pour la sélection des variables d'intérêt que la proportion de gènes sous $H1$ est importante, et que le nombre de patients augmente. Avec trop peu de patients, il se peut qu'il n'y ait aucun vrai positif sélectionné.

Il est difficile de prétendre qu'une méthode est réellement meilleure que l'autre étant donnée la largeur des intervalles de confiance. Le modèle 2 n'a pas été conservé car les résultats étaient peu différents de ceux obtenus pour le modèle ajusté retenu. Un développement plus approfondi de cette démarche n'était pas l'objectif de ce travail mais il serait intéressant de poursuivre la réflexion sur ce modèle dans de futurs travaux.

5.4.3 Généralisation des résultats obtenus

Le choix d'une méthode de sélection des gènes était indispensable ; l'objectif cependant n'était pas d'évaluer les performances de cette méthode de sélection, mais d'évaluer globalement l'influence de la phase de sélection de gènes. Nous pensons que les résultats obtenus reflètent généralement cette influence, et ne sont pas propres à l'utilisation de la méthode du gradient. Pour généraliser les résultats obtenus, il serait cependant intéressant de les confirmer en choisissant d'autres méthodes de sélection.

5.5 Remarques sur le ρ^2

La question posée dans ce travail est complexe, puisque deux phénomènes interviennent simultanément : un premier phénomène dû à la sélection des gènes qui concerne la validation des variables introduites dans le modèle, et un second phénomène dû à l'estimation des paramètres de ce modèle. Nous avons choisi d'utiliser le ρ^2 pour quantifier l'optimisme relatif aux

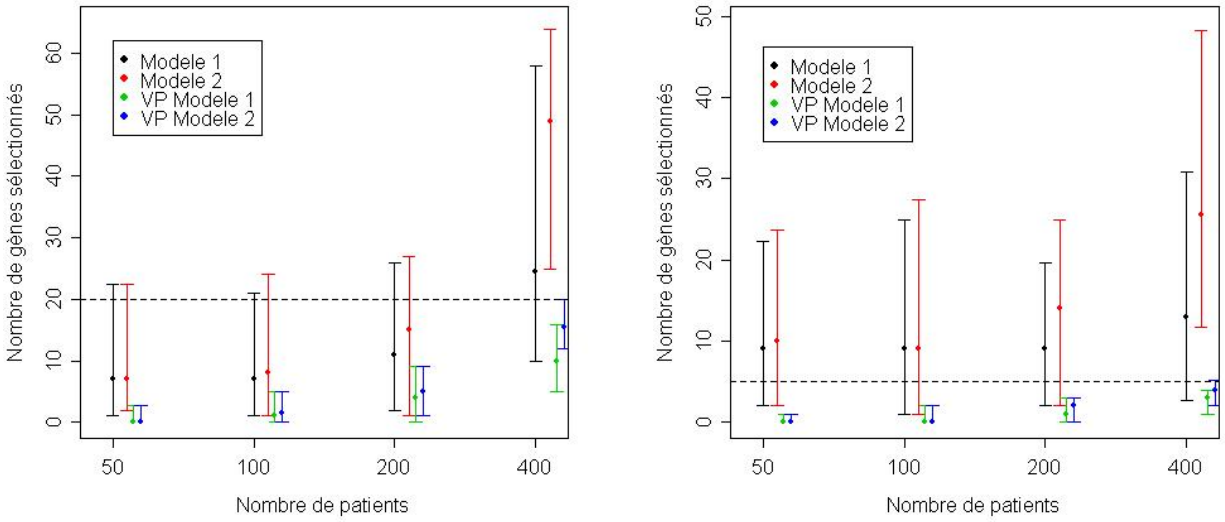


FIG. 5.9: *Évolution du nombre de gènes sélectionnés pour $p_1 = 20$ gènes d'intérêt parmi $p = 500$ (à gauche) et $p_1 = 5$ gènes d'intérêt parmi $p = 1000$ (à droite). Les segments représentent les intervalles de confiance. En noir et rouge figurent le nombre de gènes sélectionnés par les deux modèles de gradient considérés, et en vert et bleu le nombre de vrais positifs sélectionnés pour chacun d'eux.*

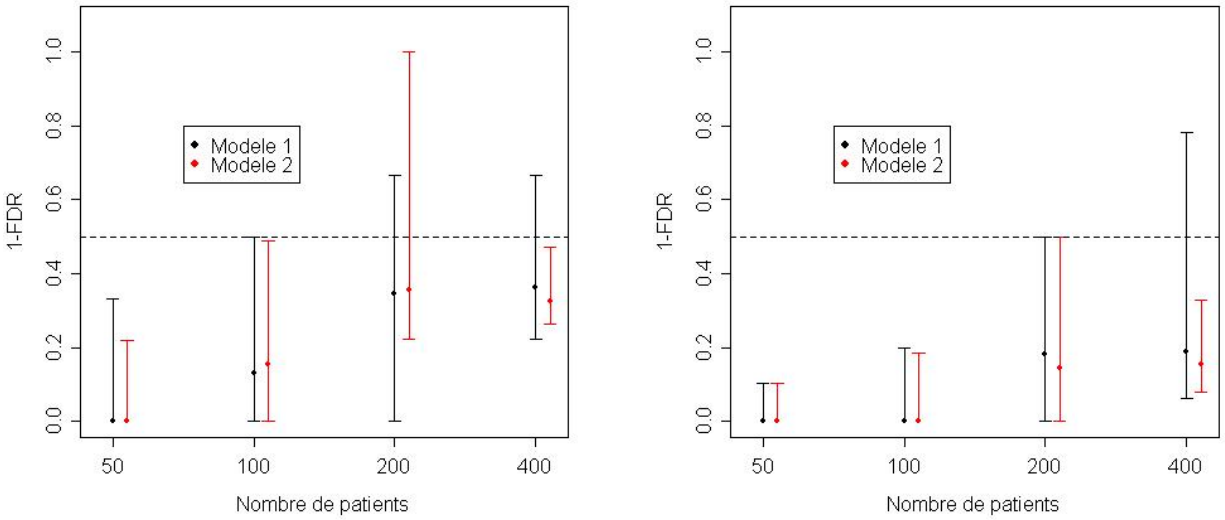


FIG. 5.10: *Évolution de (1-FDR) pour $p_1 = 20$ gènes d'intérêt parmi $p = 500$ (à gauche) et $p_1 = 5$ gènes d'intérêt parmi $p = 1000$ (à droite). Les deux couleurs correspondent aux deux modèles de gradient considérés.*

variables clinico-biologiques classiques d'une part, et aux variables transcriptomiques d'autre part. En le calculant de différentes manières (jeu test, jeu de travail, population théorique), nous avons voulu tenir compte de ces deux phénomènes. Une réflexion mérite d'être poursuivie pour quantifier de manière plus précise la part respective de chacun de ces phénomènes.

5.6 Conclusion

Lors de l'introduction de biomarqueurs clinico-biologiques classiques et de biomarqueurs génétiques, deux phénomènes interviennent : 1- la surestimation de l'effet des gènes due en partie à la phase de sélection ; 2- la sous-estimation, de l'effet des variables clinico-biologiques classiques due à l'omission de gènes d'intérêt dans le modèle.

Les biomarqueurs issus du transcriptome n'ont pas encore été validés. Ils sont souvent mis en évidence sur un unique jeu de données, et leur effet est généralisé à d'autres jeux de données. Or nos résultats montrent que la capacité prédictive de ces gènes est surestimée. Cet optimisme est d'autant plus grand que l'effectif de l'étude est faible et que le nombre de gènes étudiés est élevé. Ceci s'explique par le processus de sélection des gènes et par la trop faible puissance des études : par manque de puissance, les gènes sélectionnés sont essentiellement des faux positifs. Si la proportion de gènes d'intérêt est trop faible, l'optimisme est d'autant plus fort. C'est un constat d'autant plus gênant que le nombre de gènes d'intérêt est rarement connu à l'avance. Cependant, il peut être estimé et l'étude calibrée en fonction.

L'effet des variables clinico-biologiques classiques quant à lui n'est pas surestimé, car ces variables ont déjà été validées.

Ces remarques doivent être gardées à l'esprit lors de l'introduction des deux types de biomarqueurs dans un même modèle. L'effet des biomarqueurs classiques ne doit pas être négligé, car contrairement à celui des gènes, il n'est pas surestimé, son importance étant essentiellement masquée par l'effet observé des gènes.

Quatrième partie

Perspectives de travail

Chapitre 6

Ouverture à l'analyse du protéome

Si l'étude du transcriptome permet de quantifier le niveau d'expression des gènes d'une cellule à un moment donné, cette information n'est pas suffisante pour étudier et analyser la régulation de l'expression d'un gène dans la cellule. En effet, après la traduction interviennent des modifications post-traductionnelles, tels que l'ajout de glucides ou de lipides, le clivage et/ou le rassemblement de plusieurs chaînes polypeptidiques, qui peuvent déterminer la fonctionnalité de la protéine. Le nombre et la variété des protéines varient ainsi selon l'état et le moment de la vie de la cellule. L'ensemble des protéines dans une cellule à un moment donné constitue le protéome de la cellule. La comparaison du profil protéique d'échantillons susceptibles de présenter des différences (sain/malade, type1/type2 de tumeur, etc) ouvre donc une voie supplémentaire en clinique pour l'identification de nouveaux biomarqueurs qui vont permettre un diagnostic ou un pronostic précoce, de classer des tumeurs, constituer de nouvelles cibles thérapeutiques, etc.

La première étude clinique a été conduite par Pétricoin *et al.* [115] dans le cadre du cancer de l'ovaire. En se basant sur 50 femmes atteintes et 50 femmes indemnes de la maladie, les auteurs ont montré qu'un ensemble de 5 pics permettait de distinguer les deux groupes de femmes. Cette première étude a d'abord généré un fort enthousiasme dans la communauté scientifique...avant d'être critiquée pour son manque de rigueur [116]. Les différences de profils protéiques mises en évidence étaient en réalité dues à des artefacts techniques et non biologiques. Malgré ses limites non contestées, cette étude a ouvert le champ d'application à d'autres cancers [1] et a eu le mérite de mettre l'accent sur l'importance de la phase de pré-traitement dans l'analyse du protéome.

[1] cf Henderson et Steele [117] pour une revue des études protéomiques menées en cancérologie.

6.1 Présentation du contexte biologique

6.1.1 Acquisition des données

Le matériel biologique utilisé pour les études de protéome est classiquement le plasma ou le sérum. Deux technologies majeures sont disponibles : les gels d'électrophorèse 2D, et la spectrométrie de masse.

6.1.1.1 Electrophorèse bidimensionnelle

L'électrophorèse bidimensionnelle permet de séparer et visualiser des centaines, voire des milliers de protéines sous forme de taches sur un gel. Déposées sur un gel, les protéines contenues dans les extraits cellulaires sont séparées dans la première dimension en fonction de leur charge, puis en fonction de leur taille moléculaire dans la deuxième dimension. Les gels obtenus sont ensuite colorés puis numérisés, et l'abondance relative des protéines issues de deux échantillons différents peut être comparée sur la base des intensités de coloration des protéines séparées.

6.1.1.2 Spectrométrie de masse

La spectrométrie de masse repose également sur la séparation puis la détection des protéines présentes dans l'échantillon biologique. Après purification, l'échantillon biologique est déposé sur une lame d'acier inoxydable, prétraitée pour que la surface puisse retenir préférentiellement des classes particulières de protéines en fonction de leurs propriétés biochimiques (protéines hydrophobes, protéines anioniques ou cationiques, protéines liant des métaux, etc). Selon le type de surface utilisé, deux types de spectrométrie de masse existent : SELDI-TOF (Surface Enhanced Laser Desorption Ionisation - Time Of Flight) ou MALDI-TOF (Matrix Assisted Laser Desorption Ionisation - Time Of Flight) [118, 119]. L'échantillon biologique est mélangé avec un acide (matrice d'absorption d'énergie) qui permet sa cristallisation lorsqu'il sèche. Le cristal ainsi obtenu est placé dans un tube à vide et soumis à un rayonnement laser qui détache et ionise les protéines. Ces molécules de protéines ionisées en phase gazeuse sont soumises à un champ électrique qui produit une accélération des ions dans le tube. Enfin, un détecteur au bout du tube enregistre l'intensité et le temps de vol de chacune des molécules. Par une relation mathématique simple, à chaque temps de vol t correspond un rapport de masse sur charge m/z (mesuré en Daltons) qui va permettre d'identifier la protéine : $t = D\sqrt{m/2zV}$, où V est la tension du champ électrique appliqué, et D une constante de proportionnalité.

Un spectre est constitué de l'enregistrement du nombre d'ions (intensité) qui arrivent sur le détecteur pour un ensemble de valeurs de m/z. La figure 6.1 illustre le principe de la spectrométrie de masse. C'est cette dernière technique qui est actuellement la plus utilisée en protéomique clinique.

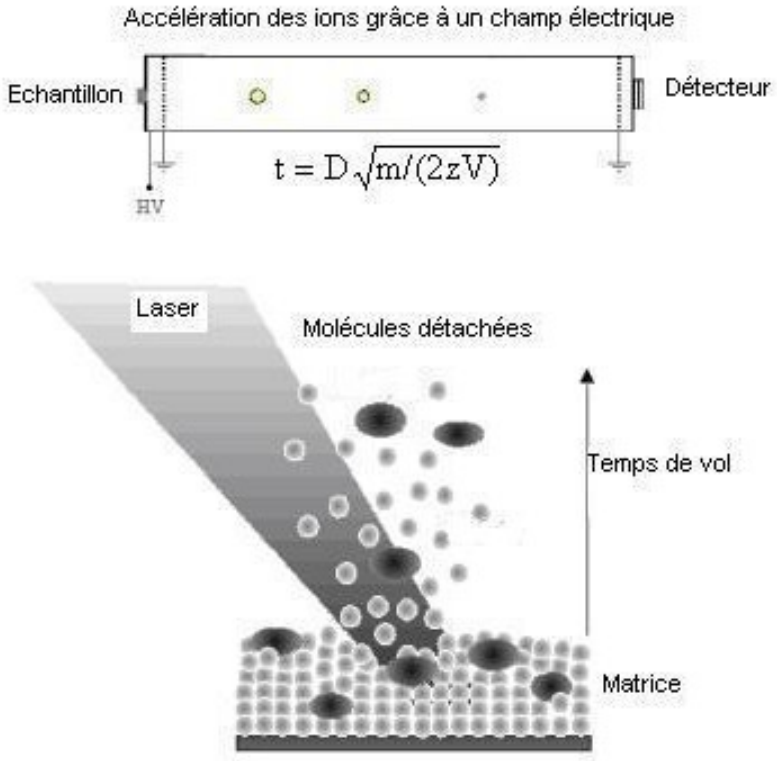


FIG. 6.1: *Principe de la mesure en spectrométrie de masse*

6.1.2 Pré-traitement des données

Comme l'analyse des biopuces, celle des spectres issues de spectrométrie de masse nécessite une phase de pré-traitement en plusieurs étapes pour soustraire de la mesure les variations qui ne sont pas des variations biologiques.

La première étape est une étape de calibration qui permet de faire correspondre le temps de vol observé à une valeur de m/z, en se basant sur un calibrant, échantillon qui contient uniquement cinq ou six protéines de masses connues.

On considère ensuite qu'un spectre est constitué par la superposition de trois composantes : le signal des pics (c'est le signal d'intérêt), un bruit de fond lisse appelé aussi ligne de base, et un bruit aléatoire de mesure. Les phases de soustraction de la ligne de base et de débruitage permettent de se rapprocher du "vrai" signal. La figure 6.2 montre l'effet de la soustraction de

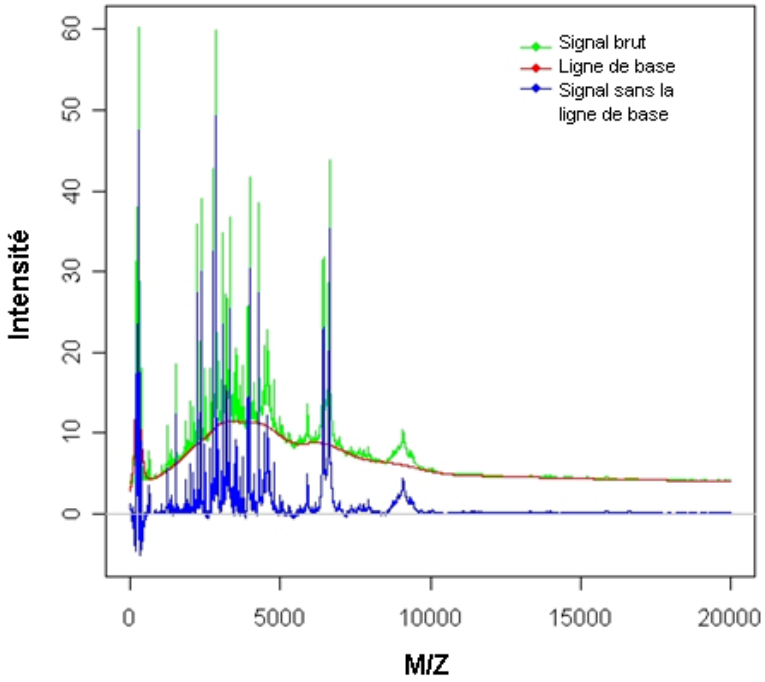


Fig. 6.2: *Visualisation sur un spectre des effets de la soustraction de la ligne de base.*

la ligne de base sur un spectre. En vert est représenté le spectre brut, en rouge la ligne de base, et en bleu le signal après soustraction de la ligne de base.

Une fois cette étape effectuée, et pour être en mesure de comparer des spectres de patients différents, les pics jugés informatifs doivent être détectés puis alignés pour faire se correspondre les pics préalablement détectés et jugés identiques d'un spectre à l'autre.

6.1.3 Traitement des données

L'analyse du protéome introduit un degré de complexité supplémentaire par rapport à l'étude du transcriptome. En effet, avec la technologie des biopuces, les variables d'intérêt potentielles sont connues *a priori* : ce sont les gènes correspondant aux sondes, dont l'emplacement sur la puce et l'identité sont connus. Le nombre de variables à étudier est donc défini. Dans l'analyse du protéome en revanche, le nombre de variables à étudier n'est pas connu *a priori*. Une étape supplémentaire d'identification des variables est nécessaire, puisque les pics qui correspondent à des protéines doivent êtres identifiés, en les différenciant des pics correspondant à du bruit.

Cette particularité des données protéomiques conduit à deux types d'approches pour leur analyse. La première consiste à travailler sur un ensemble de pics identifiés dans un certain pour-

centage des spectres, en assignant une valeur nulle aux pics non détectés dans un spectre. Les mêmes méthodes que celles utilisées pour l'étude du transcriptome peuvent alors être utilisées, avec les mêmes enjeux statistiques dus au "fléau de la dimension". Le second type d'approche permet de contourner l'étape de détection des pics en utilisant l'analyse fonctionnelle qui prend comme unité statistique non plus les pics, mais le spectre tout entier comme une fonction. La méthodologie des ondelettes est particulièrement adaptée à ce type de données [120, 121, 122].

6.2 Problématiques associées

L'implication dans l'analyse du protéome a été motivée par une collaboration avec la plateforme protéomique de Dijon, dirigée par le Docteur Patrick Ducoroy. Ce travail a été réalisé avec Catherine Mercier (ingénieur de recherche biostatisticienne) et Pascal Roy (PU-PH) du laboratoire Biostatistique-Santé, et Delphine Pecqueur (ingénieur bioinformaticienne) de la plateforme de Dijon.

L'objectif clinique est d'identifier des protéines caractérisant les patients atteints de lymphome Hodgkinien[1] à très haut risque de rechute, dans le but d'établir un pronostic précoce pour ces patients. Les données ont été recueillies par le Docteur Casasnovas du Service d'Hématologie Clinique au CHU de Dijon. Ces données ont suscité différentes questions biologiques et/ou méthodologiques.

6.2.1 Acquisition des spectres

Une étape préalable à l'analyse des spectres est la détermination du nombre de tirs nécessaires et suffisants pour récupérer l'information pertinente contenue dans le dépôt.

Pour acquérir un spectre, l'appareil somme en réalité 15 mesures acquises à des endroits différents du dépôt, chacune de ces mesures étant elle même une somme sur 100 tirs du laser à un même endroit.

L'analyse s'est basée sur deux séries de mesures concernant chacune un seul échantillon de plasma.

1. L'échantillon a été dupliqué avant purification (deux sous-protéomes issus de deux purifications du même échantillon) avec la même chimie, et déposé quatre fois pour chaque

[1] Affection cancéreuse caractérisée par une prolifération (multiplication) cellulaire anormale dans un ou plusieurs ganglions lymphatiques.

purification. Pour chaque dépôt, les spectres bruts ont été obtenus pour chacune des 15 mesures. La figure 6.3 illustre ces différents niveaux de mesures.

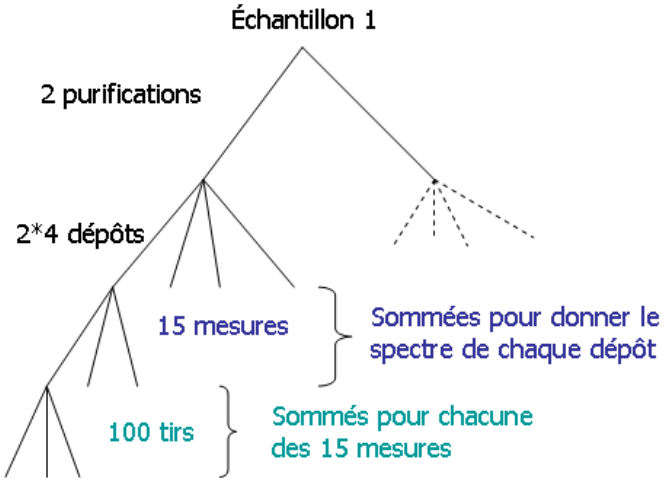


FIG. 6.3: *Illustration des différents niveaux de mesures disponibles pour le premier échantillon*

2. Sur un autre échantillon, une deuxième série de mesures a été réalisée avec 49 mesures par dépôt au lieu de 15, pour voir si les résultats obtenus sur la première série d'acquisition étaient retrouvés.

Pour évaluer la méthode de sommation, les spectres obtenus au cours des sommations intermédiaires (de 1 à 14 spectres sommés) doivent être comparés à une référence. En l'absence de gold-standard, c'est le tracé correspondant à la somme des 15 spectres qui a été pris comme tracé de référence : on a ainsi considéré que ce spectre contenait toute l'information contenue dans le dépôt.

Pour évaluer et quantifier la similarité entre deux spectres, la procédure suivante a été répétée x fois pour $n = 2, ..., 14$.

- n spectres sont tirés au sort parmi les 15 disponibles et un spectre est construit comme somme de ces n spectres. Chacun des spectres individuels est débruité suivant la méthode des ondelettes proposée par Coombes *et al.* [121]. Les spectres ne sont pas alignés avant sommation, pour reproduire ce que fait la machine au moment de l'acquisition. Par contre, une fois la somme effectuée, le spectre résultant est pré-traité et aligné par rapport au spectre de référence.
- Une distance d entre le spectre de référence et le spectre obtenu est ensuite calculée, cette distance correspondant à la somme, sur toutes les valeurs de m/z, des carrés des écarts entre le spectre sommé et la référence.

- Pour chaque valeur de n, une distribution des valeurs de cette distance a été obtenue. Tirer n spectres dans un ensemble de 15 spectres correspond à une combinaison C_{15}^{n}. Le nombre de combinaisons explose après $n = 3$. Pour que les distributions pour chaque n reposent sur le même nombre d'observations, nous avons choisi $x = 105 = C_{15}^{13} = C_{15}^{2}$, qui est le nombre de combinaisons qu'il est possible de faire en sommant 13 (ou 2) spectres ; le cas $n = 14$ a été retiré car il ne permettait que 15 combinaisons.
- La moyenne et la variance de chacune de ces distributions ont été calculées.

On s'attend à ce que la moyenne des distances tende vers zéro quand l'information contenue en sommant les n spectres tend vers celle obtenue dans le spectre de référence.

Les résultats obtenus sur les 15 mesures sont représentés sur la figure 6.4. Chaque couleur correspond à l'un des 8 dépôts. On peut considérer que l'information recueillie se stabilise après sommation d'une dizaine de spectres. En ce qui concerne les variances, elles deviennent quasiment nulles à partir de la somme de 6 spectres, ce qui signifie qu'une bonne représentation du dépôt est atteinte avec 6 spectres. On peut noter que la variance est très différente d'un dépôt à l'autre.

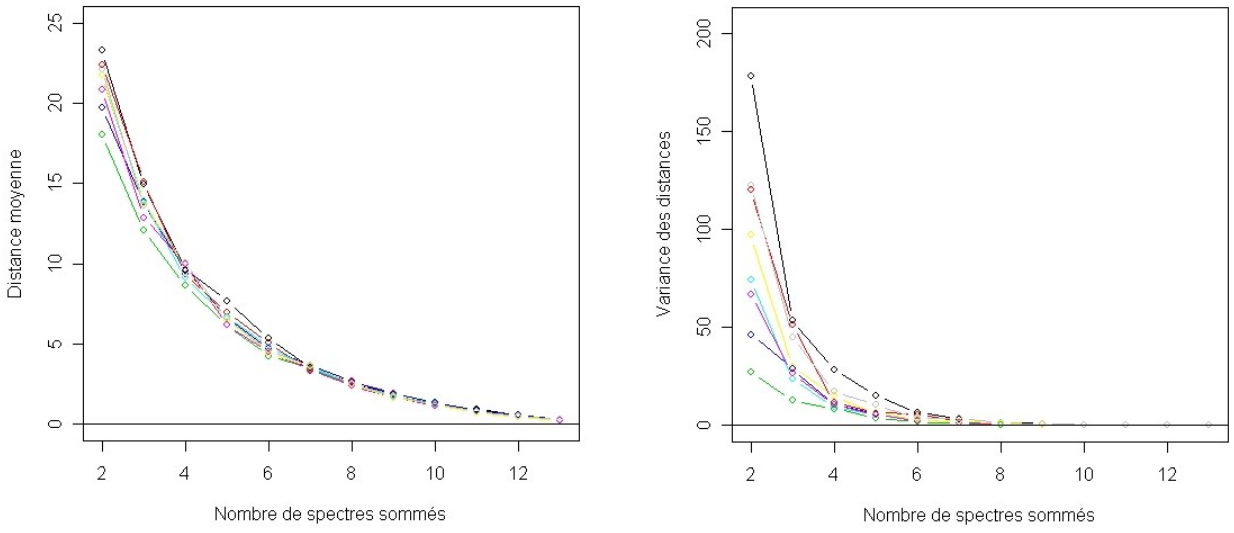


FIG. 6.4: *Évolution de la moyenne (à gauche), et de la variance (à droite) des distances entre le spectre de référence obtenu sur 15 mesures et les spectres sommés. Chaque couleur correspond à l'un des 8 dépôts.*

Les résultats avec 49 mesures (figure 6.5) confirment ceux obtenus avec 15 mesures. On observe un net décrochage des distances moyennes à partir de 7 mesures, et une stabilisation à partir d'une dizaine de spectres sommés. Les mêmes observations que ci-dessus sont également

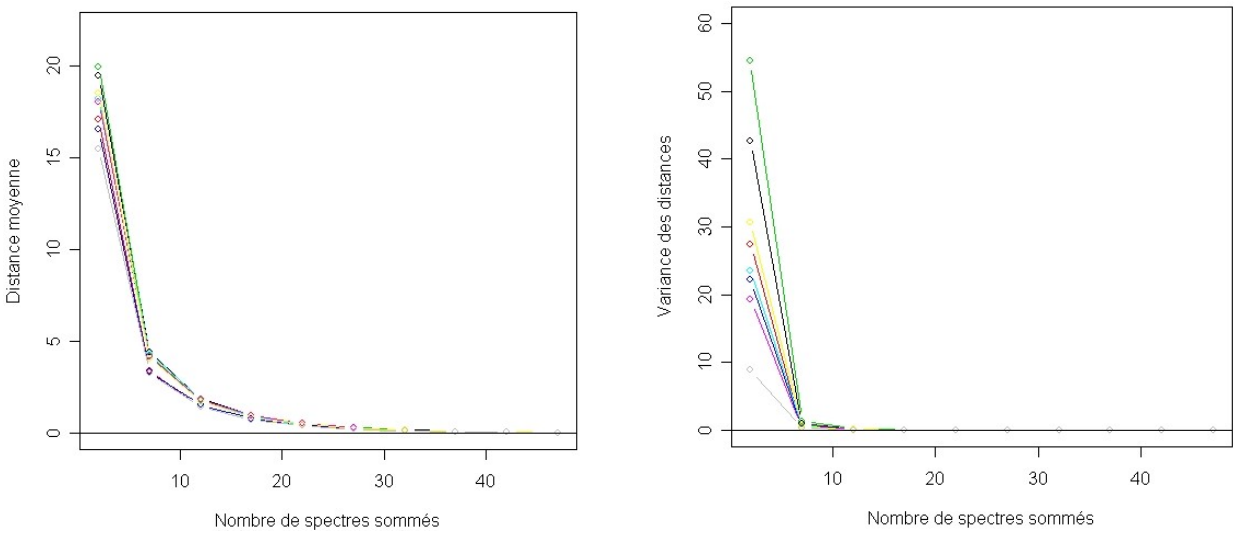


FIG. 6.5: *Évolution de la moyenne (à gauche) et de la variance (à droite) des distances entre le spectre de référence obtenu sur 49 mesures et les spectres sommés. Chaque couleur correspond à l'un des 8 dépôts.*

valables en ce qui concerne la variance des distances.

Ainsi, on peut considérer qu'une acquisition sur 10 spectres serait suffisante, plutôt que sur 15 spectres comme cela est proposé par défaut. Cette information a une conséquence directe pour le biologiste puisqu'elle permet de diminuer le temps d'acquisition des spectres.

6.2.2 Analyse de la variance

L'objectif de ce travail est de quantifier la part de variabilité inter- et intra-patients dans le but initial de simuler des spectres correspondants à une situation biologique réelle. Parmi une cohorte de patients atteints de la maladie de Hodgkin, des échantillons sont analysés chez deux groupes de 24 patients ayant ou non présentés une rechute dans un délai de 14 mois. Chaque échantillon préparé est déposé quatre fois sur la lame. L'analyse statistique concerne donc 96 spectres par groupe. La figure 6.6 illustre la manière dont sont répartis ces 96 patients.

Ce travail en cours, débuté avec Catherine Mercier, est actuellement à l'état d'ébauche, certaines pistes de travail ayant été identifiées. Pour évaluer les parts respectives de la variance nous nous orientons vers un modèle à effets mixtes. Les pics n'étant pas nécessairement les mêmes entre les groupes, les deux groupes de patients sont dans un premier temps analysés séparément. Les spectres ont été pré-traités, et les pics détectés selon un algorithme développé dans la librairie Process de R, qui sélectionne les pics après évaluation des trois critères suivants :

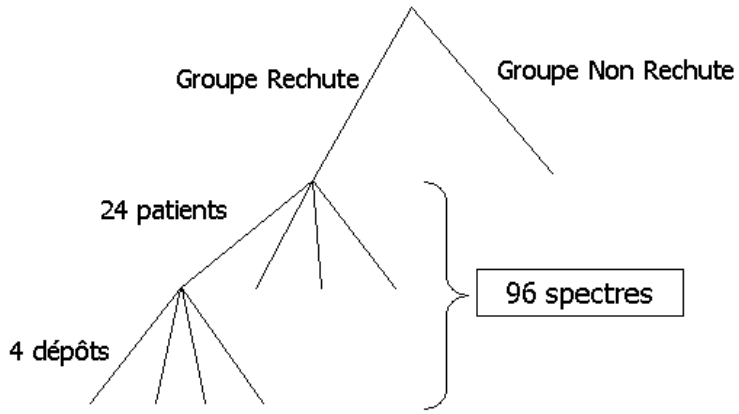


FIG. 6.6: *Répartition des 96 patients atteints de la maladie de Hodgkin.*

1- la valeur du rapport signal sur bruit autour du pic ; 2- l'intensité du pic (en dessous du seuil, l'intensité du pic est mise à zéro) ; 3- l'aire sous le pic. Parmi les pics détectés sur l'ensemble des spectres, seuls ont été conservés ceux qui étaient présents au moins dans deux spectres, conformément aux recommandations de Coombes *et al.* [121]. Une valeur nulle a été assignée aux intensités des pics absents d'un spectre.

L'influence dans l'analyse des "zéros" correspondant à l'absence d'une protéine dans un spectre est une question intéressante : cette absence peut être d'origine "biologique" (le patient ne présente pas la protéine), ou "technique" (la protéine passe sous un seuil de détection). Pour tenir compte de cette information, nous envisageons la prise en compte d'un mélange de deux distributions pour les intensités des pics : une pour les pics rares détectés chez très peu de patients (conduisant à des zéros), une pour les pics fréquents (rarement nuls).

Nous envisageons également l'introduction d'une hiérarchie croisée pour les effets des patients et des pics.

6.3 Prise en compte simultanée des différents types de biomarqueurs

Avec le développement de ces technologies à grande échelle, le clinicien est à l'heure actuelle face à trois niveaux d'information : information clinico-biologique classique, information issue du transcriptome et information issue du protéome. La deuxième partie de mon travail de thèse a permis de mettre en évidence le fait que les variables cliniques et transcriptomiques ne pouvaient

pas être considérées de la même manière dans les modèles prédictifs. La même question se pose pour les variables qui vont provenir de l'analyse du protéome. L'enjeu statistique est de voir dans quelles mesures les trois niveaux d'information peuvent être couplés dans un même modèle statistique, afin de rendre optimale la prise en charge des patients. Dans cet objectif, la réflexion sur la méthode du gradient intégrant différemment les différents types de variables pourrait être poursuivie. C'est une problématique à laquelle je souhaite répondre dans la suite de mon travail.

Bibliographie

[1] P Benkimoun. Mieux identifier les cancers du sein pour mieux les traiter. Le Monde, 13 Décembre 2006.

[2] J DeRisi, L Penland, PO Brown, ML Bittner, PS Meltzer, M Ray, YD Chen, YA Su, and JM Trent. Use of a cdna microarray to analyse gene expression patterns in human cancer. Nature Genetics, 14(4) :457–460, 1996.

[3] R Simon, M.D Radmacher, and K Dobbin. Design of studies using dna microarrays. Genetic Epidemiology, 23 :21–36, 2002.

[4] L.D Miller, P.M Long, L Wong, S Mukherjee, L.M McShane, and E.T Liu. Optimal gene expression analysis by microarrays. Cancer Cell, 2 :353–361, 2002.

[5] D.B. Allison, X. Cui, G.P Page, and M Sabripour. Microarray data analysis : from disarray to consolidation and consensus. Nature Reviews Genetics, 7(1) :55–65, 2006.

[6] T Golub, D Slonim, and P Tamayo. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. Science, 286 :531–537, 1999.

[7] A.A Alizadeh, M.B Eisen, R.E Davis, C Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, and L. M. Staudt. Distinct type of diffuse large b-cell lymphoma identified by gene expression. Nature, 403 :503–511, 2000.

[8] M.A Shipp, K.N Ross, P Tamayo, and A.P Weng. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature, 8(1) :68–74, 2002.

[9] L.J Van't Veer, H Dai, M.J Van de Vijver, Y.D He, A.A.M Hart, M Mao, H.L Peterse, K Kooy, R.M Marton, A.T Witteveen, G.J Schreiber, R.M Kerkhoven, C Roberts, P.S Linsley, E Bernards, and S.H Friend. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415 :530–536, 2002.

[10] M.J Van de Vijver, H.E Yudong, L Van't Veer, and et al. A gene expression signature as a predictor of survival in breast cancer. The New England Journal of Medicine, 347 :1999–2009, 2002.

[11] S Dudoit, Y.H Yang, M.J Callow, and T.P Speed. Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Technical Report, 2000.

[12] Y Benjamini and Y Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society,SeriesB, 57(1) :289–300, 1995.

[13] A Reiner, D Yekutieli, and Y Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures 10.1093/bioinformatics/btf877. Bioinformatics, 19(3) :368–375, 2003.

[14] J.D. Storey and R Tibshirani. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16) :9440–9445, 2003.

[15] J Aubert, A Bar-Hen, JJ Daudin, and S Robin. Determination of the differentially expressed genes in microarray experiments using local fdr. BMC Bioinformatics, 5(125), 2004.

[16] S Scheid and R Spang. Twilight ; a bioconductor package for estimating the local false discovery rate. Bioinformatics, 21(12) :2921–2922, 2005.

[17] Y Pawitan, S Michiels, S Koscielny, A Gusnanto, and A. Ploner. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics, 21(13) :3017–3024, 2005.

[18] M-L.T Lee and G.A Whitmore. Power and sample size for dna microarray studies. Statistics in Medicine, 21(23) :3543–3570, 2002.

[19] R Tibshirani. A simple method for assessing sample sizes in microarray experiments. BMC Bioinformatics, 7(106), 2006.

[20] P Muller, G Parmigiani, C Robert, and J Rousseau. Optimal sample size for multiple testing : the case of gene expression microarrays. Technical report, Johns Hopkins University, Dept. of Biostatistics, 2004.

[21] C-A Tsai, S-J Wang, D-T Chen, and James J.C. Sample size for gene expression microarray experiments. Bioinformatics, 21(8) :1502–1508, 2005.

[22] J.A Ferreira and A Zwinderman. Approximate sample size calculations with microarray data : An illustration. Statistical Applications in Genetics and Molecular Biology, 5(1, Article 25) :Available at : http ://www.bepress.com/sagmb/vol5/iss1/art25, 2006.

[23] K Dobbin and R Simon. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics, 6(1) :27–38, 2005.

[24] S-H Jung. Sample size for fdr-control in microarray data analysis. Bioinformatics, 21(14) :3097–3104, 2005.

[25] M.B Eisen, P.T Spellman, P.O Brown, and D Botstein. Cluster analysis and display of genome-wide expression patterns. Poceedings of the National Academy of Science, 95 :14863–14868, 1998.

[26] J.B MacQueen. Some methods for classification and analysis of multivariate observations. In 5-th Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297, Berkeley, 1967. University of California Press.

[27] R. Herwig, A.J. Poustka, C. Muller, C. Bull, H. Lehrach, and J. O'Brien. Large-scale clustering of cdna-fingerprinting data. Genome Research, 9(11) :1093–1105, 1999.

[28] H Hotelling. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24 :417–441 & 498–520, 1933.

[29] O Alter, P.O Brown, and D Botstein. Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences, 97(18) :10101–10106, 2000.

[30] J Landgrebe, W Wurst, and G Welzl. Permutation-validated principal components analysis of microarray data. Genome Biology, 3(4) :research0019.1 – research0019.11, 2002.

[31] J Goeman, S.A van de Geer, F de Kort, and H.C van Houwelingen. A global test for groups of genes : testing association with a clinical outcome. Bioinformatics, 20(1) :93–99, 2004.

[32] J.J Goeman, J Oosting, A-M Cleton-Jansen, J.K Anninga, and H.C van Houwelingen. Testing association of a pathway with survival using gene expression data. Bioinformatics, 21(9) :1950–1957, 2005.

[33] E Fix and J.L Hodges. Discriminatory analysis, non-parametric discrimination : consistency properties. Technical report, USAF Scholl of aviation and medicine, Randolph Field, 1951.

[34] R Tibshirani, T Hastie, B Narasimhan, and G Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences, 99(10) :6567–72, 2002.

[35] M.P.S Brown, W.N Grundy, D Lin, C.W Sugnet, T.S Furey, M Ares, and D Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Science, 97(1) :262–267, 2000.

[36] T.S Furey, N Duffy, N Cristianini, D Bednarski, M Schummer, and D Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16(10) :906–914, 2000.

[37] D.K Slonim, P Tamayo, J.P Mesirov, T.R Golub, and E.S Lander. Class prediction and discovery using gene expression data. In Fourth Annual International Conference on Computational Molecular Biology, pages 263–272, Tokyo, Japan, 2000.

[38] S Dudoit, J Fridlyand, and T Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association, 97(457) :77–87, 2002.

[39] A.C. Culhane, G Perriere, E.C. Considine, T.G. Cotter, and D.G. Higgins. Between-group analysis of microarray data. Bioinformatics, 18(12) :1600–1608, 2002.

[40] F Baty, M Bihl, G Perriere, A Culhane, and M Brutsche. Optimized between-group classification : a new jackknife-based gene selection procedure for genome-wide expression data. BMC Bioinformatics, 6(1) :239, 2005.

[41] R Diaz-Uriarte and S Alvarez de Andres. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(1) :3, 2006.

[42] L Breiman. Random forests. Machine Learning, 45(1) :5–32, 2001.

[43] L Breiman. Bagging predictors. Machine Learning, 26(2) :123–140, 1996.

[44] P Besse. Data mining 2. modélisation statistique et apprentissage, 2004.

[45] C Romualdi, S Campanaro, D Campagna, B Celegato, N Cannata, S Toppo, G Valle, and G Lanfranchi. Pattern recognition in gene expression profiling using dna array : a comparative study of different statistical methods applied to cancer classification. Human Molecular Genetics, 12(8) :823–836, 2003.

[46] A-L Boulesteix. Pls dimension reduction for classification with microarray data. Statistical Applications in Genetics and Molecular Biology, 3(1) :Article 33 ; http ://www.bepress.com/sagmb/vol3/iss1/art33, 2004.

[47] A Rosenwald, G Wright, W.C Chan, and J.M Connors. The use of molecular profiling to predict survival after chemotherapy for diffuse large b-cell lymphoma. The New England Journal of Medicine, 346(25) :1937–1947, 2002.

[48] Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences, 98 :5116–5121, 2001.

[49] R. Breitling, P. Armengaud, A Amtmann, and P. Herzyk. Rank products : a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Letters, 573(1-3) :83–92, 2004.

[50] I.B Jeffery, D. G. Higgins, and A Culhane. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics, 7 :359, 2006.

[51] S Wold, M Sjöström, and L Eriksson. Pls-regression : a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58 :109–130, 2001.

[52] D.V Nguyen and D.M Rocke. Assessing patient survival using microarray gene expression data via partial least squares proportional hazard regression. Computing Science and Statistics, 33 :376–390, 2001.

[53] D.V Nguyen and D.M Rocke. Partial least squares proportional hazard regression for application to dna microarray survival data. Bioinformatics, 18(12) :1625–1632, 2002.

[54] P Bastien. Pls cox model : Application to gene expression. In Jaromir Antoch, editor, Proceedings of 16th Symposium in Computational Statistics, pages 655–662. Physica-Verlag, Springer, 2004.

[55] H Li and J.G Gui. Partial cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics, 20(Suppl1) :i205–i215, 2004.

[56] D.V Nguyen. Partial least squares dimension reduction for microarray gene expression data with a censored response. Mathematical Biosciences, 193(1) :119–137, 2005.

[57] A-L Boulesteix and K Strimmer. Partial least squares : a versatile tool for the analysis of high-dimensional genomic data 10.1093/bib/bbl016. Briefings in Bioinformatics, 8(1) :32–44, 2007.

[58] K.C. Li. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association, 86 :316–327, 1991.

[59] Y Xia, H Tong, W Li, and L-X Zhu. An adaptive estimation of dimension reduction space. Journal of the Royal Statistical Society B, 64(3) :363–410, 2002.

[60] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc. Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics, 19(5) :563–570, 2003.

[61] F Chiaromonte and J.A Martinelli. Dimension reduction strategies for analyzing global gene expression data with a response. Mathematical Biosciences, 176 :123–144, 2001.

[62] L Li and H Li. Dimension reduction methods for microarrays with application to censored survival data. Bioinformatics, 20(18) :3406–3412, 2004.

[63] L Li. Survival prediction of diffuse large-b-cell lymphoma based on both clinical and gene expression information. Bioinformatics, 22(4) :466–471, 2006.

[64] E Bair and R Tibshirani. Semi-supervised methods to predict patient survival from gene expression data. PLOS Biology, 2(4) :511–522, 2004.

[65] P.H.C Eilers, J.M Boer, G.J.B van Ommen, and J.C van Houwelingen. Classification of microarray data with penalized logistic regression. In M.L. Bittner, Y. Chen, A.N. Dorsel, and Dougherty, editors, Proc. SPIE Vol. 4266, p. 187-198, Microarrays : Optical Technologies and Informatics, Michael L. Bittner ; Yidong Chen ; Andreas N. Dorsel ; Edward R. Dougherty ; Eds.

[66] H.C van Houwelingen, T Bruinsma, A.A.M Hart, L.J van't Veer, and L.F.A Wessels. Cross-validated cox regression on microarray gene expression data. Statistics in Medicine, 25(18) :3201–3216, 2006.

[67] P.J Verweij and H.C van Houwelingen. Cross-validation in survival analysis. Statistics in Medicine, 12(24) :2305–2314, 1993.

[68] Y Pawitan, J Bjöhle, S Wedren, K Humphreys, L Skoog, F Huang, L Amler, P Shaw, P Hall, and J Bergh. Gene expression profiling for prognosis using cox regression. Statistics in Medicine, 23 :1767–1780, 2004.

[69] R Tibshirani. A proposal for variable selection in the cox model. Technical report, University of Toronto, April 13, 1994 1994.

[70] R Tibshirani. The lasso method for variable selection in the cox model. Statistics in Medicine, 16 :385–395, 1997.

[71] B Efron, H Hastie, I Johnstone, and R Tibshirani. Least angle regression. Annals of Statistics, 32(2) :407–499, 2004.

[72] J. Gui and H. Li. Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics, 21(13) :3001–3008, 2005.

[73] M.R Segal. Microarray gene expression data with linked survival phenotypes : diffuse large-b-cell lymphoma revisited. Biostatistics, 7(2) :268–285, 2006.

[74] J.H Friedman and B.E Popescu. Gradient directed regularization. Technical report, Statistics Department, Stanford University, September 2, 2004 2004.

[75] J. Gui and H. Li. Threshold gradient descent in methods for censored data regression with applications in pharmacogenomics. In Pacific Symposium on Biocomputing, pages 17–28, Big Island of Hawaii, 2005.

[76] H Li and Y Luan. Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. Bioinformatics, 21(10) :2403–2409, 2005.

[77] H Li and Y Luan. Kernel cox regression models for linking gene expression profiles to censored survival data. In Pacific Symposium on Biocomputing, volume 8, pages 65–76, 2003.

[78] J.H Friedman. Greedy function approximation : A gradient boosting machine. Annals of Statistics, 29(5) :1189–1232, 2001.

[79] C Ambroise and GJ McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences, 99(10) :6562–6566, 2002.

[80] R Simon. Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. British Journal of Cancer, 89 :1599–1604, 2003.

[81] R Simon, M.D Radmacher, K Dobbin, and L.M McShane. Pitfalls in the use of dna microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute, 95(1) :14–18, 2003.

[82] S Michiels, S.H Koscielny, and C Hill. Prediction of cancer outcome with microarrays : a multiple random validation strategy. The Lancet, 365(9458) :488–492, 2005.

[83] L Ein-Dor, O Zuk, and E Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proceedings of the National Academy of Science, 103(15) :5923–5928, 2006.

[84] L Ein-Dor, I Kela, G Getz, D Givol, and E Domany. Outcome signature genes in breast cancer : is there a unique set ? 10.1093/bioinformatics/bth469. Bioinformatics, 21(2) :171–178, 2005.

[85] C Fan, D.S Oh, L Wessels, B Weigelt, D.S Nuyten, A.B Nobel, L.J van't Veer, and C.M Perou. Concordance among gene-expression-based predictors for breast cancer. New England Journal of Medicine, 355(6) :560–569, 2006.

[86] [http ://www.genome.wi.mit/MPR/lymphoma].

[87] R.A Irizarry, B.M Bolstad, F Collin, L.M Cope, B Hobbs, and T. P Speed. Summaries of affymetrix genechip probe level data. Nucleic Acids Research, 31(4) :e15, 2003.

[88] R Gentleman, V.J Carey, D.J Bates, B.M Bolstad, M Dettling, S Dudoit, B Ellis, L Gautier, Y Ge, J Gentry, K Hornik, T Hothorn, W Huber, S Iacus, R Irizarry, F Leisch, C Li, M Maechler, A.J Rossini, G Sawitzki, C Smith, G Smyth, L Tierney, J.Y.H Yang, and J Zhang. Bioconductor : open software development for computational biology and bioinformatics. Genome Biology, 5 :R80, 2004.

[89] B.M.I Bolstad, R.A Irizarry, M. Astrand, and T.P Speed. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics, 19(2) :185–193, 2003.

[90] J.D Emerson and D.C Hoaglin. Analysis of two-way tables by medians. In D.C Hoaglin, F Mosteller, and J.W. Tukey, editors, Understanding robust and exploratory data analysis., volume 27, pages 166–206. John Wiley & Sons, Inc, New York, 1983.

[91] D Singh, P.G Febbo, K Ross, D.G Jackson, J Manola, C Ladd, P Tamayo, A.A Renshaw, A.V D'Amico, and J.P Richie. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 1(2) :203–209, 2002.

[92] [http ://www.genome.wi.mit/MPR/prostate].

[93] U Alon, N Barkai, DA Notterman, K Gish, S Ybarra, D Mack, and AJ. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences U S A., 96(12) :6745–6750, 1999.

[94] T Golub. golubEsets : exprSets for Golub leukemia data. R package version 1.0.

[95] E Tian, F Zhan, R Walker, E Rasmussen, Y Ma, B Barlogie, and JD Shaughnessy. The role of the wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma. The New England Journal of Medicine, 25(349) :2483–2494, 2003.

[96] [http ://www.ncbi.nlm.nih.gov/geo].

[97] S Chiaretti, X Li, R Gentleman, A Vitale, M Vignetti, F Mandelli, J Ritz, and R Foa. Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival 10.1182/blood-2003-09-3243. Blood, 103(7) :2771–2778, 2004.

[98] Z Wu, R.A Irizarry, R Gentleman, F Martinez-Murillo, and F Spencer. A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association, 99 :909, 2004.

[99] S Dolédec and D Chessel. Rythmes saisonniers et composantes stationnelles en milieu aquatique i- description d'un plan d'observations complet par projection de variables. Acta Ňcologica, Ňcologia Generalis, 8(3) :403–426, 1987.

[100] Y Guo, H Hastie, and R Tibshirani. Regularized discriminant analysis and its application in microarray. Technical report, Department of Health research and Policy, May 5, 2004 2004.

[101] D. V Nguyen. On partial least squares dimension reduction for microarray-based classification :.a simulation study. Computational statistics & data analysis, 46 :407–425, 2004.

[102] H Li. Censored data regression in high-dimension and low-sample size settings for genomic applications. UPenn Biostatistics Working Papers, Working Paper 9, 2006.

[103] C Truntzer, C Mercier, J Estève, C Gautier, and P Roy. Importance of data structure in comparing two dimension reduction methods for classification of microarray gene expression data. BMC Bioinformatics, 8(90), 2007.

[104] D.V Nguyen and D.M Rocke. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics, 18(1) :39–50, 2002.

[105] A-L Boulesteix. Reader's reaction to "dimension reduction for classification with gene expression microarray data" by dai et al. Statistical Applications in Genetics and Molecular Biology, 5(1) :Article 16 ;Available at : http ://www.bepress.com/sagmb/vol5/iss1/art16, 2006.

[106] J.J Dai, L Lieu, and D (2006) Rocke. Dimension reduction for classification with gene expression microarray data. Statistical Applications in Genetics and Molecular Biology, 5(1) :Article 6. Available at : http ://www.bepress.com/sagmb/vol5/iss1/art6, 2006.

[107] L Lebart, A Morineau, and M Piron. Statistique exploratoire multidimensionnelle. Paris, 1995.

[108] M. Tenenhaus and F.W Young. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis ans other methods for quantifying categorical multivariate data. Psychometrika, 50 :91–119, 1985.

[109] Y Escoufier. The duality diagramm : a means of better practical applications. In : Development in numerical ecology. Serie G .Springer Verlag. Legendre, P. & Legendre, L., Berlin, 1987.

[110] M Barker and W Rayens. Partial least squares for discrimination. Journal of Chemometrics, 17 :166–173, 2003.

[111] C Truntzer, D Maucort-Boulch, and P Roy. Model optimism and comparative contribution of clinical and tanscriptomic variables. Submitted, 2007.

[112] JT. Kent and J O'Quigley. Measures of dependence for censored survival data. Biometrika, 75 :525–534, 1988.

[113] JT. Kent. Information gain and a general measure of correlation. Biometrika, 70 :163–174, 1983.

[114] C Chastang, D Byar, and S Piantadosi. A quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models. Statistics in Medicine, 7(12) :1243–1255, 1988.

[115] E.F Petricoin, A.M Ardekani, and B.A Hitt. Use of proteomic patterns in serum to identify ovarian cancer. The Lancet, 359 :572–577, 2002.

[116] K.A Baggerly, J.S Morris, and K.R Coombes. Reproducibility of seldi-tof protein patterns in serum : comparing datasets from different experiments. Bioinformatics, 20(5) :777–785, 2004.

[117] N.A Henderson and R.J. Steele. Seldi-tof proteomic analysis and cancer detection. Surgeon, 3(6) :383–390, 2005.

[118] T.W Hutchens and T.T Yip. New desorption strategies for the mass-spectrometric analysis of macromolecules. Rapid Communications in Mass Spectrometry, 7(576-580), 1993.

[119] M. Merchant and S.R. Weinberger. Recent advancements in surfaceenhanced laser desorption/ionization-time of flight-mass spectroscopy. Electrophoresis, 21 :1164–1177, 2000.

[120] J.S. Morris and R.J Caroll. Wavelet-based functional mixed models. Journal of the Royal Statistical Society, Series B, 68(2) :179–199, 2006.

[121] K.R Coombes, Jr Fritsche, H.A, C Clarke, J Chen, K.A Baggerly, J.S Morris, L Xiao, M-C Hung, and H.M Kuerer. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. Clinical Chemistry, 49(10) :1615–1623, 2003.

[122] P Du, W.A. Kibbe, and S.M. Lin. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. Bioinformatics, 22(17) :2059–2065, 2006.

Cinquième partie

Annexes

Annexe A

Premier article - publié

BMC Bioinformatics

BioMed Central

Research article

Open Access

Importance of data structure in comparing two dimension reduction methods for classification of microarray gene expression data

Caroline Truntzer*[1], Catherine Mercier[1], Jacques Estève[1], Christian Gautier[2] and Pascal Roy[1]

Address: [1]CNRS, UMR 5558 – Equipe Biostatistique Santé, Villeurbanne, F-69100, France, Université Claude Bernard Lyon 1, Laboratoire Biostatistique Santé – UMR 5558, Villeurbanne, F-69100, France, Hospices Civils de Lyon, Service de Biostatistique, Lyon, F-69003, France and [2]Université Claude Bernard – Lyon 1, Laboratoire de Biométrie et de Biologie Evolutive – UMR CNRS 5558, Villeurbanne, F-69100, France

Email: Caroline Truntzer* - caroline.truntzer@chu-lyon.fr; Catherine Mercier - catherine.mercier@chu-lyon.fr; Jacques Estève - jacques.esteve@chu-lyon.fr; Christian Gautier - cgautier@biomserv.univ-lyon1.fr; Pascal Roy - pascal.roy@chu-lyon.fr

* Corresponding author

Published: 13 March 2007

BMC Bioinformatics 2007, **8**:90 doi:10.1186/1471-2105-8-90

Received: 6 October 2006
Accepted: 13 March 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/90

Abstract

Background: With the advance of microarray technology, several methods for gene classification and prognosis have been already designed. However, under various denominations, some of these methods have similar approaches. This study evaluates the influence of gene expression variance structure on the performance of methods that describe the relationship between gene expression levels and a given phenotype through projection of data onto discriminant axes.

Results: We compared Between-Group Analysis and Discriminant Analysis (with prior dimension reduction through Partial Least Squares or Principal Components Analysis). A geometric approach showed that these two methods are strongly related, but differ in the way they handle data structure. Yet, data structure helps understanding the predictive efficiency of these methods. Three main structure situations may be identified. When the clusters of points are clearly split, both methods perform equally well. When the clusters superpose, both methods fail to give interesting predictions. In intermediate situations, the configuration of the clusters of points has to be handled by the projection to improve prediction. For this, we recommend Discriminant Analysis. Besides, an innovative way of simulation generated the three main structures by modelling different partitions of the whole variance into within-group and between-group variances. These simulated datasets were used in complement to some well-known public datasets to investigate the methods behaviour in a large diversity of structure situations. To examine the structure of a dataset before analysis and preselect an a priori appropriate method for its analysis, we proposed a two-graph preliminary visualization tool: plotting patients on the Between-Group Analysis discriminant axis (x-axis) and on the first and the second within-group Principal Components Analysis component (y-axis), respectively.

Conclusion: Discriminant Analysis outperformed Between-Group Analysis because it allows for the dataset structure. An a priori knowledge of that structure may guide the choice of the analysis method. Simulated datasets with known properties are valuable to assess and compare the performance of analysis methods, then implementation on real datasets checks and validates the results. Thus, we warn against the use of unchallenging datasets for method comparison, such as the Golub dataset, because their structure is such that any method would be efficient.

Background

In cancer research, microarray technology offers a new tool for diagnosis of specific tumors or prognosis of survival. However, in microarray experiments, there are more variables (genes) than samples (patients); if not taken into account, this dimension problem leads to trivial results with no statistical identifiability or biological significance.

Among the methods proposed to overcome this problem, some look for discriminant axes that best separate distinct groups of patients according to specific characteristics. These discriminant axes define a new space whose dimension is lower than that of the original gene space. The discriminant axes are constructed as linear combinations of genes; that is, each gene contributes to the construction of the axes through a coefficient (weight) that depends on its importance in discriminating the groups. Then, for prediction purposes, new patients may be projected in this lower space and assigned to the nearest group. This article focuses on three types of discriminant analysis widely used for prediction purposes: Principal Component Analysis (PCA) followed by Discriminant Analysis (DA), Partial Least Squares followed by DA, and Between-Group Analysis (BGA).

DA is proposed to define discriminant axes [1,2]. One concern in DA is that it is limited by "high dimensionality" and requires a preliminary dimension reduction step. The classical approach to dimension reduction is PCA [3] where components are such that they maximize the gene expression variability across samples. Another approach coming from chemometrics, the PLS method [4-8], selects the components that maximize the covariance between gene expression and phenotype response. To circumvent this preliminary step within the context of microarray data analysis, Culhane *et al*. [9] proposed the Between-Group Analysis [10], because it can be directly used even when the number of variables exceeds the number of samples.

A few recent publications were dedicated to comparisons between projection methods within the context of microarray data analysis. Nguyen and Rocke compared PCA and PLS as prior procedures to logistic discrimination or quadratic discriminant analysis [11]. Boulesteix studied PLS+DA in more detail [12]. Dai *et al*. proposed a new comparison between PCA and PLS extended to a comparison with the Sliced Inverse Regression (SIR) dimension reduction method [13] as prior to logistic discrimination. At the same time, Jeffery *et al*. [14] pointed out that the variance structure of the dataset mostly influences the efficiency and comparison of feature selection methods. No similar work has been done to see whether the structure of the variance of a given dataset may impact the efficiency of the above-cited projection methods. Thus, bioinformaticians may encountered difficulties in choosing the most adapted method for a given dataset.

To solve these difficulties, we found it of major importance to extend the previous comparison studies by a detailed look at the properties of DA -with previous PCA or PLS- and BGA, to understand how some a priori knowledge of the dataset structure may help choosing the most appropriate method.

To achieve this goal, we used both simulated and public well-know datasets in a complementary approach. As to simulated datasets, the article presents a novel simulation process to model various data structures, which leads to different partitions of the whole variance into within-group and between-group variances. A special attention is given to the case where one discriminant axis separates two groups; e.g., whenever a given phenotype classifies the patients into two groups (for example, tumor vs. non-tumor patients). The overall results are discussed to provide appropriate recommendations for more efficient microarray analysis.

Methods

General analysis scheme

BGA and DA are based on the same principle: finding one discriminant linear combination of genes that defines a direction in $\mathbb{R}^p$ (gene space) along which the between-group variance is maximized. The methodology of multidimensional analysis provides an appropriate framework [15]. Consider a $(n * p)$ data array X that gives for each n patients on rows the values of p gene expression levels. Each column, the expression of one gene, is a vector of $\mathbb{R}^n$ and each row, the set of gene expression for one patient of the population, is a vector in $\mathbb{R}^p$. The aim was to detect a relationship between patients and genes and find a subspace that provides the best adjustment of the scatter plot. This adjustment requires the definition of a metric in $\mathbb{R}^p$, given by a (p, p) positive symmetric matrix Q that defines a scalar product and distances in $\mathbb{R}^p$.

Introducing information about groups is necessary to find a subspace in which the between-group variance is maximum. This may be reached through introduction of a matrix of indicators Y, which enables group identification to be incorporated in a new matrix Z. BGA and DA follow the same general analysis scheme using this matrix Z and specific choices for Q [16].

Definition of Z

Let the (n, k) matrix Y, containing k class indicators, define a partition of the n patients. To maximize the between-group variance, columns of X are projected on the subspace defined by the columns of Y. This projection is obtained through the projection operator P_Y defined as: P_Y

$= Y({}^tYY)^{-1}({}^tY)$. Projecting patients on a class of k indicators is equivalent to computing the mean expression of each variable in class k. P_YX is a (n, p) matrix where the variables for each patient are replaced by the corresponding means of the class he belongs to. Actually, the rank of this matrix is k - 1. With this choice of $Z = P_YX$, maximizing the variance of a linear combination of Z is equivalent to maximizing the between-group variance of X. BGA and DA may be seen as a PCA of the mean matrix, each having its own metric in $\mathbb{R}^p$. As said above, BGA does not require a preliminary dimension reduction before projecting patients on the discriminant axis. However, DA requires dimension reduction, which leads first to express patients of X in a lower subspace. X_{red} contains the patients coordinates in this reduced space. Z_{red} is then a (n, p) matrix where variables for each patient in the previously reduced space are replaced by the corresponding means of the class he belongs to.

Two methods are classically proposed to reduce dimension: normed PCA and PLS. They yield components that are linear combinations of genes considered as the new variables to analyze by DA [11]. Each of those components includes all the initial variables weighted according to their contribution to the effect caught by the component. PCA aims at finding components that maximize the projected variance of the data. In contrast, PLS looks directly for components associated with the phenotype. Only a subset of the first components is sufficient to catch most of the data variance or covariance. The optimal number of components was chosen by cross-validation, as described by Boulesteix in the case of PLS+DA [12].

Choice of Q

Once Z chosen, BGA and DA derive from two distinct choices for Q. In BGA, $Q = I_p$ where I_p is the (p, p) identity matrix. In DA, the metric $Q = ({}^tXX)^{-1}$, so the metric involves the total variance-covariance matrix for all patients whatever their group. Another metric could be the mean of the intra-group variances. It corresponds to the so-called Linear Discriminant Analysis. The total variance being the sum of within-group and between-group variances, there is a direct relationship between the two methods. Whatever the metric, the assumption is that variance-covariance matrices are similar in all groups. Moreover, in both cases, the metric involves an inversion of $({}^tXX)$, which requires not too strongly correlated variables. This is not typically the case in microarray studies due to the huge number of variables, which calls for dimension reduction.

Statistical solution

The general analysis applies to any pair (Z, Q). In BGA, the pair is $(Z, I_p) = (P_YX, I_p)$; in DA, it is $(Z_{red}, ({}^tX_{red}X_{red})^{-1})$. The general scheme aims at finding linear combinations $Z\alpha$ maximizing $\|Z\alpha\|_{I_n}$, where α is a (p, r) matrix. Those linear combinations define a subspace in which the variance of Z is maximum. The single solution is given by singular value decomposition of the matrix $Q({}^tZ)Z$. This matrix can always be diagonalized and has p eigenvalues with r non-zero ones λ_i, $i = 1...r$. The r corresponding eigenvectors maximize $\|Z\alpha\|_{I_n}$ under Q^{-1}-orthonormality constraint; they are defined in $\mathbb{R}^p$, and called principal factors. Columns of α contain these eigenvectors. By definition, the α_i are Q^{-1}-normed. With this construction, linear combinations are uncorrelated.

In the particular case discussed here, where Z corresponds to a mean table for two groups, there is only one discriminant axis, so $r = 1$. In the general case of k groups, $r = k$ - 1.

Performance estimator

BGA and DA were compared using their predictive performances; i.e., the proportion of correctly classified patients.

The phenotype of a new patient was predicted according to its position on the discriminant axis relative to the threshold defined as:

$$\frac{\bar{X}_{G1}SD_{G2} + \bar{X}_{G2}SD_{G1}}{SD_{G1} + SD_{G2}} \qquad (1)$$

In Equation (1), $\bar{X}_{G1}$, $\bar{X}_{G2}$, SD_{G1} and SD_{G2} are respectively the means and standard deviations of the two groups. This threshold was proposed by Culhane *et al*. [9] for BGA and used here also for DA. It allows taking into account the accuracy of the assignment, a greater weight being given to the less scattered group.

Following the idea of Boulesteix [12], Leave-k-Out Cross-Validation was used to obtain the proportion of correctly classified patients. In each loop, the dataset was randomly split so that $k = 1/3$ of the samples were left out and the model derived using the 2/3 samples was applied to predict the class of the remaining samples. This operation was repeated fifty times and a mean misclassification proportion computed. With DA, the selection of the number of components was included in the cross-validation process. The mean misclassification proportion was determined for each number of components used as variables. Finally, the number of components kept was the one for which the misclassification proportion over the fifty runs was minimal.

The variability of the performance estimator (PE) was measured somewhat differently with simulated and real datasets. With simulated datasets and a given set of parameters, the standard deviation of the PE was computed over the fifty simulated datasets. This informs about the variability stemming from the whole process used for PLS+DA, PCA+DA, or BGA. The standard deviation of the PE over the fifty cross-validation runs was computed for each real dataset and for the optimal number of components. This shows to which extent the choice of the split that led to build the training sets may influence the proportion of well-classified samples of the test set, with the same number of components kept.

Implementation of methods

All computations were performed using R programming language. The R code that enables to perform simulations is available as additional file [see Additional file 1]. To perform BGA, we used the made4 library [17]. To perform DA with prior PLS or PCA, we relied on the plsgenomics library [18].

Gene expression datasets

DLBCL

This dataset contains 7,129 expression levels on 58 patients with Diffuse Large B-Cell Lymphoma (DLBCL) [19]. After preprocessing and use of a filter method, only 6,149 expression levels were kept. These patients are divided into two subgroups depending on the 5-year survival outcome: 32 "cured" patients and 26 "fatal/refractory" patients. The data are available as .CEL files from the Broad Institute website [20]. The gene expression values were called using the Robust Multichip Average method and data were quantile normalized using the Bioconductor package *affy* [21].

Prostate

This dataset provides 102 samples: 50 without and 52 with prostate tumors [22]. The data are available as .CEL files from the Broad Institute website [23]. The gene expression values were obtained as above.

ALL

This dataset includes 125 patients with Acute Lymphoblastic Leukemia [24]: 24 patients with and 101 without multidrug resistance (MDR). The pre-processed data are available in the *ALL* library in Bioconductor [21].

Leukaemia

This well-known dataset includes expression data on 7,129 genes from 72 tumor-mRNA samples [25]. These acute leukaemia samples belong to two different subtypes of leukaemia: 27 samples categorized as ALL (Acute Lymphoblastic Leukemia) and 45 categorized as AML (Acute Myeloid Leukemia), which is the phenotype of interest. Data are available in the *golubEsets* library in Bioconductor [21]. The data were processed by making the min expression value 100 and the max expression value 16,000. The log_2 of the data was then used.

Results

The datasets used herein are either artificial data obtained by an original simulation process or the above-cited two-class public datasets.

Simulated datasets

Simulation process

Simulations were performed as a first step to understand the influence of data structure on the results with DA and BGA. An original simulation process was carried out to evaluate the extent to which the above procedures were able to retrieve the structure of a simple two-component problem. We modeled different partitions of the whole variance into within-group and between-group variances using three parameters: i) the variance-covariance structure of each group; ii) the length of the vector joining the barycenters of the two groups; and iii) the direction of this vector, toward a high or a low within-group variance. These three parameters result in several relative positions and eccentricities of the scatter plots in the two-component space.

The simulations started with the generation, in the component space, of two groups with known within-group variances. The maximum dimension of this component space is n, the number of patients of the datasets. The between-group difference was expressed in the two-component space. In this space, variables were drawn from a bivariate normal distribution $N(\mu, \Sigma)$ where Σ is a $(2 * 2)$ diagonal matrix with elements σ_1 and σ_2. μ depended on the distance *dist* between the barycenters of the scatter plots.

Thus, *dist* allowed controlling the between-group structure. The chosen ratio σ_1/σ_2 reflects eccentricity: the higher it is, the higher is the eccentricity of the scatterplots; so, this ratio allowed controlling the within-group structure. The line joining the barycenters of the groups and the first component axis forms an angle α. Figure 1 shows the geometric meaning of these parameters. The n - 2 dimensions left correspond to noise.

Next, patients were expressed in the $\mathbb{R}^n$ gene space. For this, gene axes were derived from the component axes through a chosen rotation, which masks more or less the between-group structure present in the two-component space.

The p - n genes left are random linear combinations of these n genes.

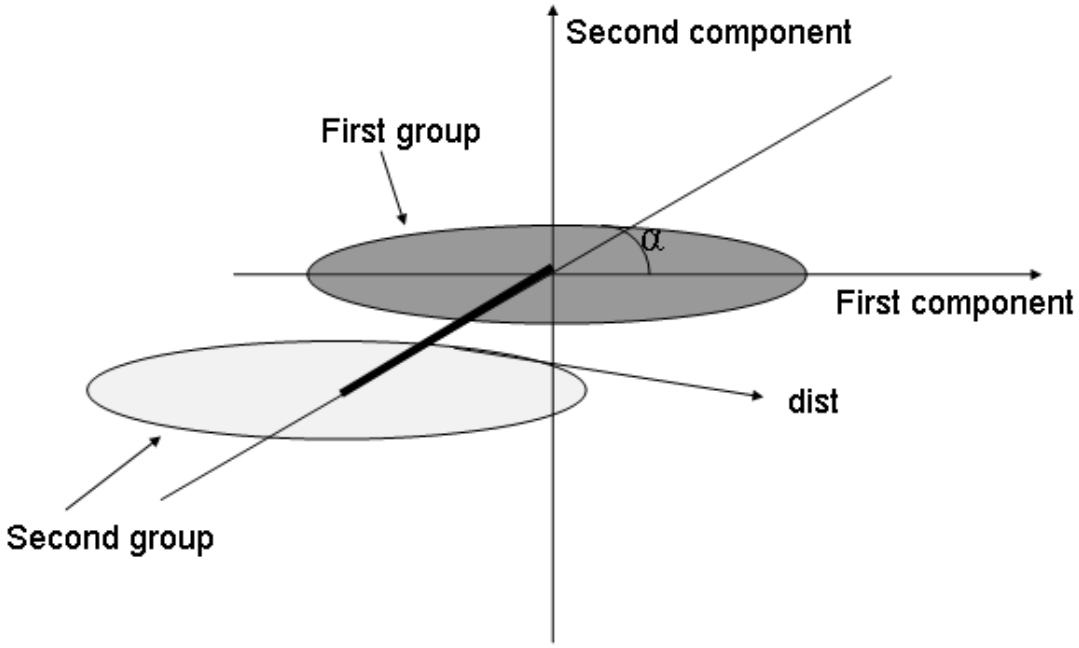


Figure 1
View of the component space relative to the simulations. The cluster of points of each of the two groups was plotted in the two-component space. The scatter plots barycenters are distant by *dist*. The direction of the between-group variance draws an angle α with the first component.

The effect of dist and α

Table 1 shows prediction results for distances *dist* equal to 1, 3, and 5. The observed differences between DA and BGA did not depend on the previous dimension reduction method, PCA or PLS. However, the number of components kept was always greater (or equal) with PCA than with PLS and, in some cases, this advantaged PCA+DA as seen for *dist* = 1 and $\alpha = \pi/4$, for example.

Whatever the method used and the value of α, prediction was better as *dist* increased; that is, when the clusters of points were the more distant. Moreover, the more distant the barycenters were, the less the difference between DA and BGA was.

Then, for a given distance, prediction results depended on the value of α. The results with DA or BGA were the closest for $\alpha = \pi/2$ and $\alpha = 0$: both inefficient with $\alpha = 0$ and both very efficient with $\alpha = \pi/2$. This corresponded to situations where the between-group direction was simulated on the first or second component axis. For intermediate angles, both methods were less good predictors, with nevertheless an advantage for DA.

The effect of eccentricity

Table 2 shows prediction results for several α and eccentricities defined by *ratio* = σ_1/σ_2. A *ratio* of 1 corresponds to a spherical cluster of points. As expected, the higher the ratio was, the more advantageous was DA over BGA. Moreover, except for $\alpha = 0$, both methods performed generally better when eccentricity was high. With non-spherical scatter plots, the best prediction was achieved with $\alpha = \pi/2$; that is, when the between-group direction was perpendicular to the within-group direction. When the ratio decreased, DA and BGA got closer, the greatest difference being with *ratio* = 10.

Table 3 shows the results when the main components of the group variances were extremely different; that is, when the directions of the principal component of the two clusters of points were perpendicular. In that case, DA and BGA had similar results whatever α. Note that PCA was less efficient; in fact, the between-group part was low in the whole variance structure.

As a general remark, it may be noted that the standard deviation of the performance estimator over the fifty sim-

Table 1: Proportion of well-classified patients according to dist, the distance between the barycenters of the two groups

	$\alpha = \pi/2$	$\alpha = \pi/3$	$\alpha = \pi/4$	$\alpha = \pi/6$	$\alpha = 0$
dist = 1					
PLS+DA	0.69(0.05)//2	0.69(0.06)//2	0.64(0.06)//2	0.60(0.06)//2	0.59(0.05)//1
PCA+DA	0.69(0.04)//3	0.70(0.05)//2	0.66(0.06)//3	0.60(0.05)//3	0.58(0.06)//1
BGA	0.63(0.05)	0.61(0.06)	0.55(0.06)	0.57(0.05)	0.58(0.05)
dist = 3					
PLS+DA	0.93(0.04)//2	0.91(0.03)//2	0.86(0.03)//2	0.71(0.03)//2	0.69(0.06)//1
PCA+DA	0.94(0.04)//2	0.91(0.03)//2	0.85(0.04)//3	0.71(0.04)//3	0.69(0.05)//2
BGA	0.90(0.04)	0.79(0.03)	0.73(0.03)	0.70(0.03)	0.67(0.06)
dist = 5					
PLS+DA	0.98(0.04)//2	0.97(0.01)//2	0.97(0.02)//2	0.84(0.03)//2	0.79(0.04)//1
PCA+DA	0.99(0.01)//3	0.98(0.01)//2	0.97(0.02)//2	0.83(0.03)//2	0.79(0.04)//2
BGA	0.91(0.04)	0.91(0.01)	0.86(0.03)	0.82(0.03)	0.79(0.04)

Mean (Standard deviation)//Median of the optimal number of components over fifty datasets simulated with eccentricity such that *ratio* = 10. PLS: Partial Least Squares – PCA: Principal Components Analysis – DA: Discriminant Analysis.

ulated datasets was low whatever the variance partition examined.

Comparable results of simulations were obtained when differences were expressed in two- or three-component spaces.

Real datasets

The public datasets were chosen to cover the main situations encountered in practice.

To begin the analysis of a new dataset, we suggest to first have a look at its structure to visualize the relative role of the within-group and between-group variances for distinguishing the two groups of patients. For this, we propose two graphs obtained by plotting patients on the BGA discriminant axis (x-axis) and on the first and the second within-group PCA component (y-axis), respectively. The greatest part of the between-group variance is given by the most differential genes, while the other genes tend to mask this between-group structure. For this prior examination of the data structure, we used only the fifty genes with the highest t-test statistics.

Figures 2 to 5 show the plots that correspond to each dataset. In the case of the DLBCL dataset (Figure 2), the clusters of points were not discrete; the cluster relative to the cured patients was even found within the "fatal/refractory" cluster. This suggests that the dataset has no obvious between-group structure. Moreover, the main components of the variances in each group were very different.

In the case of the prostate dataset (Figure 3), the distinction between non-tumor and tumor samples was found along both between-groups and the first within-group directions.

In the case of the ALL dataset (Figure 4), the distinction between patients with or without multidrug resistance (MDR) was found along the first within-group direction.

Table 2: Proportion of well-classified patients according to ratio, which reflects eccentricity

	$\alpha = \pi/2$	$\alpha = \pi/3$	$\alpha = \pi/4$	$\alpha = \pi/6$	$\alpha = 0$
ratio = 10					
PLS+DA	0.82(0.05)//2	0.81(0.03)//2	0.76(0.03)//2	0.71(0.04)//2	0.59(0.04)//2
PCA+DA	0.85(0.05)//3	0.81(0.04)//3	0.77(0.05)//3	0.73(0.05)//3	0.59(0.04)//3
BGA	0.76(0.05)	0.75(0.04)	0.66(0.03)	0.67(0.04)	0.58(0.04)
ratio = 2					
PLS+DA	0.68(0.05)//1	0.65(0.04)//1	0.65(0.05)//1	0.65(0.05)//2	0.63(0.05)//1
PCA+DA	0.69(0.05)//3	0.65(0.04)//3	0.67(0.04)//2	0.65(0.04)//2	0.63(0.04)//2
BGA	0.67(0.06)	0.62(0.04)	0.64(0.05)	0.65(.04)	0.62(0.05)
ratio = 1					
PLS+DA	0.60(0.05)//2	0.62(0.05)//1	0.64(0.05)//2	0.62(0.05)//1	0.61(0.05)//1
PCA+DA	0.63(0.04)//3	0.63(0.04)//2	0.63(0.05)//2	0.64(0.05)//2	0.63(0.05)//2
BGA	0.61(0.05)	0.62(0.05)	0.61(0.05)	0.61(0.05)	0.60(0.05)

Mean (Standard deviation)//Median of the optimal number of components over over fifty datasets simulated with a distance between the barycenters *dist* = 2. PLS: Partial Least Squares – PCA: Principal Components Analysis – DA: Discriminant Analysis.

Table 3: Proportion of well-classified patients with a high eccentricity (ratio = 10) in one group and a low eccentricity (ratio = 0.1) in the other group

	DLBCL	Prostate	ALL	Leukaemia
PLS+DA	0.51(0.14)//12	0.97(0.06)//10	0.73(0.05)//10	0.97(0.03)//1
PCA+DA	0.49(0.09)//13	0.96(0.07)//9	0.57(0.08)//1	0.95(0.04)//5
BGA	0.43(0.10)	0.70(0.09)	0.60(0.06)	0.98(0.03)

Mean (Standard deviation)//Median of the optimal number of components over fifty datasets simulated with *dist* = 2. PLS: Partial Least Squares – PCA: Principal Components Analysis – DA: Discriminant Analysis.

At last, in the case of the leukaemia dataset (Figure 5), the barycenters were only separated by the between-group direction. This indicates that the between-group direction was perpendicular to the within-group direction.

So, these four datasets reflect various structures of variance; these structures may be associated to simulated datasets to see how their main characteristics explain the predictive behaviour of the methods. Table 4 shows the proportion of well-classified patients obtained over the fifty cross-validation runs with the optimal number of components. The standard deviation of the performance estimator over the fifty cross-validation runs was low. This standard deviation shows the variability of the performance estimator between cross-validation runs. Here, it indicated that the way of splitting patients into training and test sets within each run did not affect the results.

As expected in the light of the structure visualization, the proportions of well-classified samples for the DLBCL were low whatever the method used, BGA being the less efficient. In fact, DA needed 12 PLS components or 13 PCA components to optimize prediction while, with only one component, BGA is not able to catch more information given by the within-group structure. This corresponded in the simulated datasets to a low value of *dist*.

As to the prostate dataset, the plots led to compare this dataset to the case where α is intermediate between 0 and $\pi/2$. Thus, we could foretell that the results would be improved in comparison with those of the DLBCL dataset, and that DA will be more advantageous. Indeed, this was confirmed with the proportions of well-classified samples: DA was more efficient in predicting non-tumor or tumor samples. It seemed that the high number of components kept for the first dimension reduction allowed getting more information than a single projection in BGA.

The ALL dataset corresponded to simulating α near to 0; none of the methods was really adapted to such a configuration. Actually, no methods was sufficiently efficient. PCA as first dimension reduction method was not able to catch information. On the contrary, with 10 PLS components, DA overcame BGA.

As to the leukaemia dataset, it recalled the simulated case with $\alpha = \pi/2$, which is the one that allowed the best results. This was confirmed in Table 4, where the three methods were particularly efficient in distinguishing ALL and AML patients. The prediction results obtained with BGA and DA were very similar. With dimension reduction, one PLS component and five PCA components were needed to optimize prediction. The results with PCA suggested that the between-group variance took the largest part of the total variance.

Further figures are provided as additional files showing the structure of other well-known datasets: DLBCL vs FL [see Additional file 2], Colon (normal vs tumor samples) [see Additional file 3], Myeloma (With vs without lytic lesions) [see Additional file 4], ALL1 (B-Cell vs T-Cell origin) [see Additional file 5], ALL2 (Relapse vs no relapse) [see Additional file 6], ALL3 (With vs without t(9;22) translocation) [see Additional file 7]. The corresponding proportions of well-classified patients obtained over the fifty cross-validation runs with the optimal number of components are provided in additional file 8 [see Additional file 8].

Discussion

Results from both simulated and real datasets showed that the structure of a dataset influences to a large extent the efficiency of the methods that use projection on discriminant axes.

In testing a new method, simulated and real datasets play complementary roles. Simulation of data with known properties is useful to study the influence of the dataset characteristics and the performance of a given method, and could be considered as a practical guide to understand results from real situations. For choosing an analysis method to discriminate two groups of patients, we think it is necessary to have a prior examination of the structure of the data to analyze. This will enable an informed choice between the available methods.

We propose here a new simulation approach that allows exploring known structures with control through several parameters. Nguyen [26] proposed to simulate datasets to

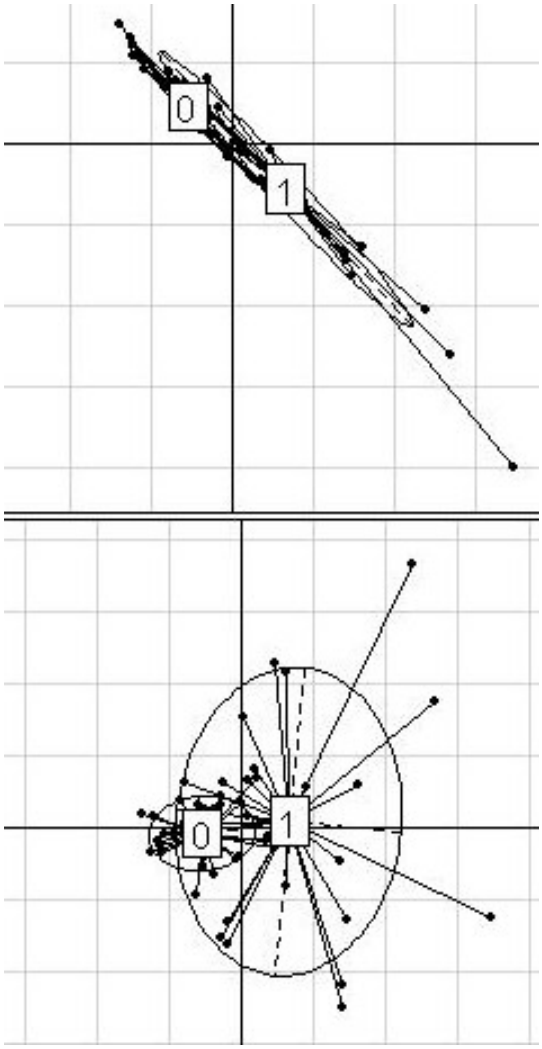


Figure 2
DLBCL dataset. Projection of the 58 patients from the DLBCL dataset (32 "cured" and 26 "fatal/refractory") on the discriminant axis obtained with BGA (x-axis), along their coordinates on the first (on the top) and the second (on the bottom) within-group PCA component (y-axis), respectively. For a better legibility, the groups were labeled 0 (for "cured" patients) and 1 (for "fatal/refractory" patients). Only the 50 most differential genes among 6149 were used for these graphs.

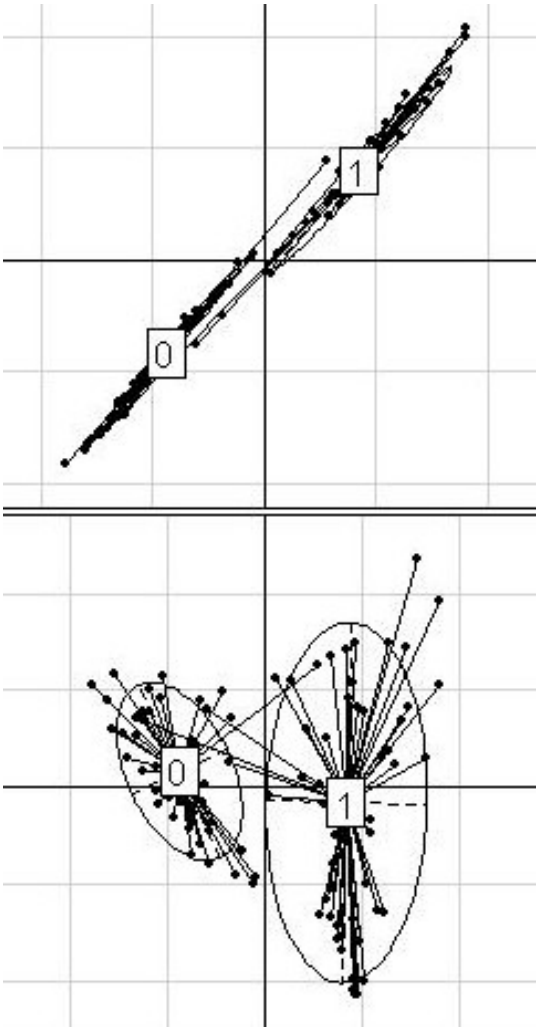


Figure 3
Prostate dataset. Projection of the 102 patients from the prostate dataset (50 without and 52 with tumor) on the discriminant axis obtained with BGA (x-axis), along their coordinates on the first (on the top) and the second (on the bottom) within-group PCA component (y-axis), respectively. For a better legibility, the groups were labeled 0 (for non-tumor prostate samples) and 1 (for tumor prostate samples). Only the 50 most differential genes among 12625 were used for these graphs.

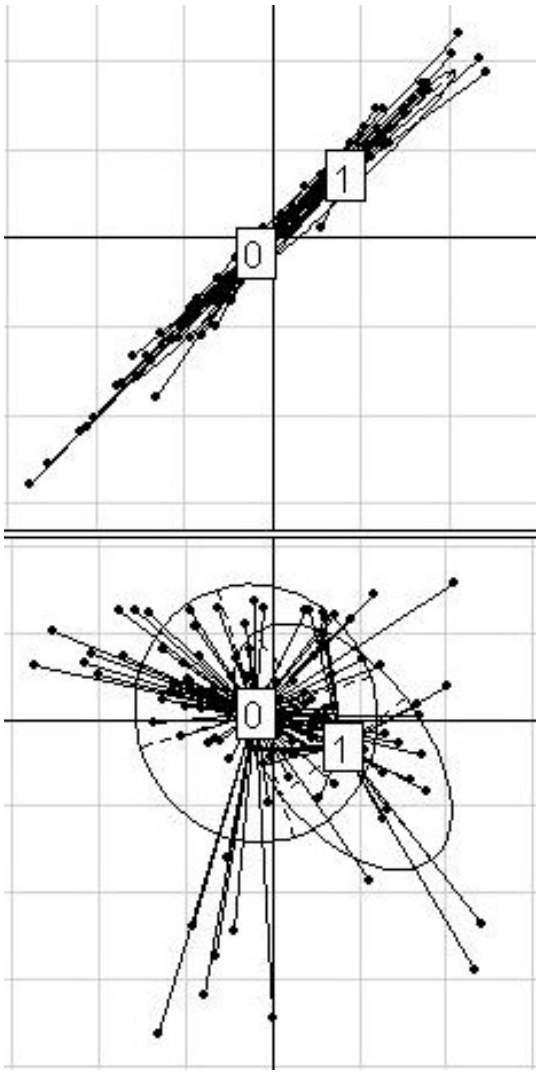


Figure 4
ALL dataset. Projection of the 125 patients from the ALL dataset (24 with and 101 without Multi Drug Resistance -MDR-) on the discriminant axis obtained with BGA (x-axis), along their coordinates on the first (on the top) and the second (on the bottom) within-group PCA component (y-axis), respectively. For a better legibility, the groups were labeled 0 (for patients with MDR) and 1 (for patients without MDR). Only the 50 most differential genes among 12625 were used for these graphs.

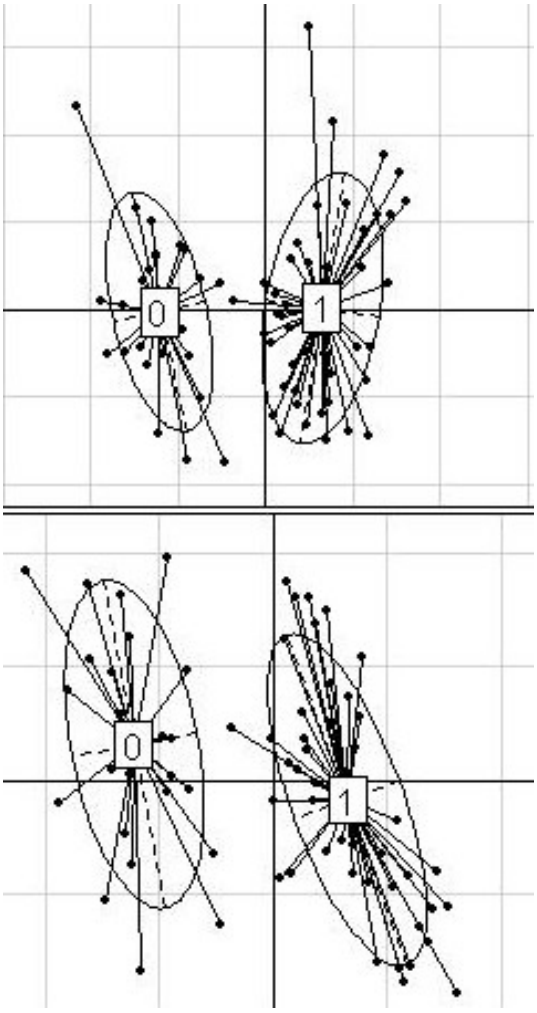


Figure 5
Leukemia dataset. Projection of the 72 patients from the leukaemia dataset (25 Acute Lymphoblastic Leukemia -ALL- and 47 Acute Myeloide Leukemia -AML-) on the discriminant axis obtained with BGA (x-axis), along their coordinates on the first (on the top) and the second (on the bottom) within-group PCA component (y-axis), respectively. For a better legibility, the groups were labeled 0 (for AML patients) and 1 (for ALL patients). Only the 50 most differential genes among 7129 were used for these graphs.

Table 4: Proportion of well-classified patients with real datasets

	PLS+DA	PCA+DA	BGA
DLBCL	0.51(0.14)//12	0.49(0.09)//13	0.43(0.10)
Prostate	0.97(0.06)//10	0.96(0.07)//9	0.70(0.09)
ALL	0.73(0.05)//10	0.57(0.08)//1	0.60(0.06)
Leukaemia	0.97(0.03)//1	0.95(0.04)//5	0.98(0.03)

Mean (Standard deviation) over the fifty cross-validation runs for the optimal number of component (indicated after //). Results were obtained with the DLBCL, the prostate, the ALL, and the leukaemia datasets. PLS: Partial Least Squares – PCA: Principal Components Analysis – DA: Discriminant Analysis.

compare the performance of PCA and PLS as prior procedure before logistic discrimination. However, his method of simulation did not allow a discussion on the influence of the data structure. Our simulations allow generating different structures of different degrees of complexity and assessing the impact of three parameters: the distance between the clusters, the eccentricity of these clusters, and their relative positions in a two-dimensional component space. The major source of complexity in real microarray datasets is the existence of regulation networks. In our simulations, this may be described by a component with a very large variance; that is, a large eccentricity. This corresponds usually to a common effect on all the genes. A high variance on one component corresponds also to a cluster of highly correlated genes. Whether a network of genes exists or not would determine the relative importance of the other components with respect to the first one. Nevertheless, we are aware that our simulations have limits. Therefore, a compromise has to be found between the uncontrolled nature of real datasets and the controlled nature of simulated datasets as research tools. This will be the object of future works.

The use of real datasets to prove the superiority of any method should be considered with caution. For example, the leukaemia dataset from Golub, very often used to demonstrate the efficiency of a new method, may not be used for that purpose because of its very strong between-group structure. This structure is such that we expect the groups to be distinguished whatever the method used (e.g., BGA that simply joins the barycenters of the groups). We believe that, in such situations, the good performance of a particular method does not only inform on its ability to discriminate between groups. If the structure of the dataset had been previously examined before its analysis, for example with the graphical tool we propose, this dataset would not have been chosen to validate new prediction methods. Thus, bioinformaticians should be cautious in choosing the datasets to use for method comparisons. The proposed visualization tool helps in choosing the dataset, by having an idea of its structure. The prostate or ALL datasets for example may be appropriate for that purpose.

Besides, the structure of a given dataset may depend on the type of disease. In diagnosis, some pathophysiological entities may be already clearly identified; if their origin is a metabolic activation, they will induce different processes that will be easy to distinguish (e.g., ALL vs. AML). However, differentiating patients with or without multidrug resistance may be even more difficult because no pathophysiological entities are involved. In prognosis, distinguishing good from bad prognosis patients would be more difficult because they often share the same pathophysiological characteristics.

Three main configurations of the data structure may be identified. When the clusters of points are quite distinct the between-group difference is so obvious that the within-group structure will have no impact; BGA and DA will give good prediction results. The simple method that consists in drawing an axis between the barycenters is sufficient. In fact, the way of projecting patients on the discriminant axis does not come into consideration. On the opposite, there are situations in which both methods are inappropriate. This corresponds to superposed clusters of points obtained in plotting the within-group versus the between-group coordinates. In other situations, we believe that DA is more advantageous than BGA because it allows taking into account the partition of the total variance into between and within variances. However, in case the variances of the two groups are not the same, the total variance will not reflect the variance in each group, so there will be no advantage of favoring DA over BGA. Moreover, keeping more than one component in the first dimension reduction step using PLS or PCA is a way to capture more information than the single projection in BGA, particularly with PLS. This is illustrated with the ALL dataset; by keeping ten PLS components, DA outperforms BGA to a large extent (respectively 0.97% and 0.70% of well-classified patients). These observations illustrate the fact that the first PLS component and the BGA discriminant axis are identical. This was demonstrated by Barker and Rayens [27], and by Boulesteix [12]. Thus, using PLS with one component followed by DA gives a final component that is collinear to that of PLS alone, and also to the BGA axis. This is illustrated with the leukaemia dataset,

where PLS+DA and BGA give equivalent results (respectively 0.97% and 0.98% of well-classified patients). However, in simulations, PLS+DA seemed to yield, on average, slightly better results than BGA. In fact, due to random sampling, some simulated datasets needed more than one component to optimize prediction because dimensions other than those simulated may be informative by chance alone. Note that in case of a spherical cluster of points, a second PLS component will not capture more information than the first one and both methods will be equally efficient.

Overall, DA becomes advantageous when the structure of the variance is such that the way of projecting patients on the discriminant axis needs to come into consideration. This leads to conclude that DA is the most suitable method; it provides better or at least equivalent results in a diversity of datasets because it ensures that the within-group variance will be taken into account, when relevant. The diversity of real datasets encountered confirms the fact that, unlike DA, BGA is unable to deal with too complex data structures. The only advantage of BGA is its ease of use and interpretation: a single projection enables to go from the original variable space to a one-dimension axis on which inter-group variance is maximum.

This axis is also a direct linear combination of genes where a high coefficient means that the gene is important to classify the patients into one of the groups. With DA, the samples are first expressed in a component space, which makes interpretation more difficult.

BGA and DA used with more than two groups provide k - 1 discriminant axes, which enables each of the k groups to be separated from the k - 1 others. By plotting these groups in successive two-dimensional graphs, the structure assessment described here may be applied to each of the two-dimension spaces so obtained.

Conclusion

We have established here that the two methods -BGA and DA with prior PCA or PLS- are based on very similar approaches. Efficient use of these projection methods requires some a priori knowledge of the structure of the clusters of points. We found that three main structure situations may be identified. When the clusters of points are clearly split, both methods will perform equally well and it becomes futile to prove the superiority of one method over the other using datasets previously shown of simple structure. When the clusters of points superpose, both methods will fail to yield interesting predictions. In such a case, there is no linear way to separate groups, leading to the use of non linear methods. In intermediate situations, the structure of the clusters of points has to be taken into account by the projection to improve prediction, which imposes the use of DA. So, we recommend the use of Discriminant Analysis to take into account more diverse dataset structures.

Authors' contributions

CT wrote the computer code for simulations, carried out the analysis, analyzed the results and drafted the manuscript. JE and PR contributed to simulations design, result interpretation, and contributed with CM and GC to write the manuscript. All authors read and approved the final manuscript.

Additional material

Additional file 1

***R codes used to generate simulated datasets**. This simulation process generates several datasets structures by modelling different partitions of the whole variance into within-group and between-group variances.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-90-S1.pdf]

Additional file 2

***DLBCL vs FL dataset**. Projection of 58 patients with Diffuse Large B-Cell Lymphoma and 19 patients with Follicular Lymphoma on the discriminant axis obtained with BGA (x-axis), along their coordinates on the first (on the top) and the second (on the bottom) within-group PCA component (y-axis), respectively. For a better legibility, the groups were labeled 0 (for FL-patients) and 1 (for DLBCL-patients). Only the 50 most differential genes among 7129 were used for these graphs. The data are available from the Broad Institute website [20].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-90-S2.jpeg]

Additional file 3

***Colon dataset**. Projection of 22 normal controls and 40 tumor samples on the discriminant axis obtained with BGA (x-axis), along their coordinates on the first (on the top) and the second (on the bottom) within-group PCA component (y-axis), respectively. For a better legibility, the groups were labeled 0 (normal controls) and 1 (for tumor samples). Only the 50 most differential genes among 2000 were used for these graphs. The data are available in the* ColonCA *library in Bioconductor [21].*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-90-S3.jpeg]

Additional file 4

***Myeloma dataset**. Projection of 36 patients with and 137 patients without lytic lesions on the discriminant axis obtained with BGA (x-axis), along their coordinates on the first (on the top) and the second (on the bottom) within-group PCA component (y-axis), respectively. For a better legibility, the groups were labeled 0 (lytic lesions) and 1 (without lytic lesions). Only the 50 most differential genes among 12625 were used for these graphs. Data can be download from Gene Expression Omnibus [28] (accession number GDS531).*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-90-S4.jpeg]

Additional file 5

***ALL1 dataset**. Projection of 95 Acute Lymphoblastic Leukaemia (ALL) patients with B-Cell and 33 with T-Cell origin on the discriminant axis obtained with BGA (x-axis), along their coordinates on the first (on the top) and the second (on the bottom) within-group PCA component (y-axis), respectively. For a better legibility, the groups were labeled 0 (B-Cell) and 1 (T-Cell). Only the 50 most differential genes among 12625 were used for these graphs. The data are available in the* GOstats *library in Bioconductor [21].*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-90-S5.jpeg]

Additional file 6

***ALL2 dataset**. Projection of 65 ALL patients that did and 35 that did not relapse on the discriminant axis obtained with BGA (x-axis), along their coordinates on the first (on the top) and the second (on the bottom) within-group PCA component (y-axis), respectively. For a better legibility, the groups were labeled 0 (no relapse) and 1 (relapse). Only the 50 most differential genes among 12625 were used for these graphs. The data are available in the* GOstats *library in Bioconductor [21].*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-90-S6.jpeg]

Additional file 7

***ALL3 dataset**. Projection of 26 ALL-patients with and 67 ALL-patients without the t(9;22) translocation on the discriminant axis obtained with BGA (x-axis), along their coordinates on the first (on the top) and the second (on the bottom) within-group PCA component (y-axis), respectively. For a better legibility, the groups were labeled 0 (without t(9;22)) and 1 (with t(9;22)). Only the 50 most differential genes among 12625 were used for these graphs. The data are available in the* GOstats *library in Bioconductor [21].*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-90-S7.jpeg]

Additional file 8

***Proportion of well-classified patients for complementary two-class real datasets**. Mean (Standard Deviation) over the fifty cross-validation runs for the optimal number of component (indicated after //). The table shows results for the following datasets: DLBCL vs FL, Colon, Myeloma, ALL1, ALL2, and ALL3.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-8-90-S8.pdf]

Acknowledgements

We wish to thank Daniel Chessel for his valuable comments and Jean Iwaz for editing the manuscript. The work was supported by a grant from the French National Cancer League given to CT. This work was also a part of a clinical research project, Pharmacogenoscan, supported by the Canceropole Lyon Auvergne Rhone-Alpes (CLARA).

References

1. Fisher R: **The use of multiple measurements in taxonomic problems.** *Ann of Eugenics* 1936, **7:**179-188.
2. Mahalanobis P: **On the generalized distance in statistics.** *Proc Nat Acad Sci India* 1936, **12:**49-55.
3. Hotelling H: **Analysis of a complex of statistical variables into principal components.** *J Educ Psychol* 1933, **24:**417-441. & 498–520
4. Garthwaite P: **An interpretation of Partial Least Squares.** *J Am Stat Assoc* 1994, **89(425):**122-127.
5. DeJong S: **SIMPLS: an alternative approach to partial least squares regression.** *Chemometr Intell Lab Syst* 1993, **18(3):**251-263.
6. Martens H, Naes T: *Multivariate calibration* New York: Wiley; 1989.
7. Stone M, Brooks R: **Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principla components regression.** *J R Statist Soc B* 1990, **52:**237-269.
8. Frank I, Friedman J: **A statistical view of some chemometrics regression tools.** *Technometrics* 1993, **35:**109-148.
9. Culhane A, Perriere G, Considine E, Cotter T, Higgins D: **Between-group analysis of microarray data.** *Bioinformatics* 2002, **18(12):**1600-1608.
10. Doledec S, Chessel D: **Rythmes saisonniers et composantes stationnelles en milieu aquatique.** *Acta Oecologica Oecologia Generalis* 1987, **8:**403-426.
11. Nguyen D, Rocke D: **Tumor classification by partial least squares using microarray gene expression data.** *Bioinformatics* 2002, **18:**39-50.
12. Boulesteix A: **PLS Dimension Reduction for Classification with Microarray Data.** *Stat Appl Genet & Mol Biol* 2004, **3:**Article 33 [http://www.bepress.com/sagmb/vol3/iss1/art33].
13. Dai J, Lieu L, Rocke D: **Dimension Reduction for Classification with Gene Expression Microarray Data.** *Stat Appl Genet & Mol Biol* 2006, **5:**Article 6 [http://www.bepress.com/sagmb/vol5/iss1/art6].
14. Jeffery I, Higgins D, Culhane A: **Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data.** *BMC Bioinformatics* 2006, **7:**359.
15. Lebart L, Morineau A, Piron M: *Statistique exploratoire multidimensionnelle* Paris: Dunod; 1995.
16. Escoufier Y: **The duality diagramm: a means of better practical applications.** In *Development in numerical ecology* Edited by: Serie G. Springer Verlag, Berlin: Legendre, P. & Legendre, L; 1987.
17. Culhane A, Thioulouse J, Perriere G, Higgins D: **MADE4: An R package for Multivariate Analysis of Gene Expression Data.** *Bioinformatics* 2005, **21(11):**2789-90.
18. Boulesteix AL, Strimmer K: *plsgenomics: PLS analyses for genomics* 2005 [http://cran.r-project.org/src/contrib/Descriptions/plsgenomics.html]. [R package version 1.0]
19. Shipp M, Ross K, Tamayo P, Weng A: **Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning.** *Nature* 2002, **8:**68-74.
20. [http://www-genome.wi.mit.edu/mpr/lymphoma].
21. [http://www.bioconductor.org].
22. Singh D, Febbo P, Ross K, Jackson D, Manola J, Ladd C, Tamayo P, Renshaw A, D'Amico A, Richie J: **Gene Expression Correlates of Clinical Prostate Cancer.** *Cancer Cell* 2002, **1:**203-209.
23. [http://www-genome.wi.mit.edu/mpr/prostate].
24. Chiaretti S, Li X, Gentleman R, Vitale A, Vignetti M, Mandelli F, Ritz J, Foa R: **Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival.** *Blood* 2004, **103:**2771-2778.
25. Golub T, Slonim D, Tamayo P: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
26. Nguyen D: **On partial least squares dimension reduction for microarray-based classification:a simulation study.** *Comput Stat Data Anal* 2004, **46:**407-425.
27. Barker M, Rayens W: **Partial least squares for discrimination.** *J Chemom* 2003, **17:**166-173.
28. [http://www.ncbi.nlm.nih.gov/geo].

Annexe B

Second article - soumis

STATISTICS IN MEDICINE
Statist. Med. 2000; **00**:1–6 *Prepared using* simauth.cls *[Version: 2002/09/18 v1.11]*

Model optimism and comparative contribution of clinical and tanscriptomic variables

Caroline Truntzer [1,*,†], Delphine Boulch [1], John O'Quigley [2] and Pascal Roy [1]

[1] *CNRS, UMR 5558 - Equipe Biostatistique Sante, Villeurbanne;*
Universite Claude Bernard Lyon 1, Laboratoire Biostatistique Sante - UMR 5558, Villeurbanne;
Hospices Civils de Lyon, Service de Biostatistique, Lyon
[2] *Institut Curie, Paris*

SUMMARY

In cancer research, most clinical variables have been already investigated and are now well established. Simultaneously, transcriptomic variables have raised two problems: restricting their number and validating their significance. Thus, their contribution to prognosis is currently thought to be overestimated. The main issue addressed here is to evaluate to which extent the optimism of the current transcriptomic models may lead to overestimate the contribution of transcriptomic variables to survival prognosis. To achieve this goal, Cox proportional hazards models that adjust differently for clinical and transcriptomic variables were built. The relevance of the clinical variables being established, these were not submitted to selection. As to genes, they were selected using the Threshold Gradient Descent method. Optimism and contribution to prognosis of clinical and transcriptomic variables, were compared through simulations and use of the Kent and O'Quigley R measure of dependence. We showed that the optimism relative to the clinical variables was low because these are no more submitted to selection of relevant variables. On the contrary, for genes, the selection process introduced a great optimism that increased when the proportion of genes of interest decreased. However, this optimism can be decreased by increasing the number of samples.

KEY WORDS: optimism; transcriptomic variables; survival models; explained randomness

1. INTRODUCTION

For a long time classical clinical variables were used as prognostic markers in the field of cancer research. Indeed, many strong clinical determinants that explain most of the prognosis have been already identified. Nevertheless, cancer still appears as a puzzle and a better understanding of its determinants is still needed to improve treatment. Thus, cancer research has headed for new technologies, especially microarrays, and survival analysis methods have been extended to take into account the potential information stemming from microarrays; this transcriptomic information would overcome the clinical one. Clinician would like now to

*Correspondence to: Caroline Truntzer, Service de Biostatistique, 162 Avenue Lacassagne, 69003 Lyon, France
†E-mail: caroline.truntzer@chu-lyon.fr

Received November 2006

combine in same models both types of variables, that is genes and classical clinical variables, to increase the precision in predicting the cancer prognosis. But the nature of transcriptomic variables is completely different than it is for clinical variables and has to be taken account in the statistical models.

Specific clinical variables have been validated in many previous studies that involved several patient samples. Thus, most of these clinical variables are no longer in the selection process (e.g. the estrogen receptor status in breast cancer or the international prognostic index in lymphoma). In contrast, the selection step is still needed for genes. Microarray analyses are rather recent and various ill-known issues are still debated. First, fewer studies have been conducted on transcriptomic than on clinical variables; thus, there are less datasets available to repeat the analyses and validate the relationships. In most cases, genes selected in a single study are assumed to have a general prognostic value. This selection is presented as a bench mark for the disease without external validation studies on new datasets. Second, whenever available, those datasets are rather small compared to the number of genes under study. Considering the high number of variables and the relatively low number of observations, microarray data lead easily to a high number of false-positive variables. By chance alone, many genes may be found significantly associated with the outcome while the majority of them would not be linked to prognosis.

A few recent publications enlightened some additional issues related to the selection process in microarray analysis. Ein-Dor et al. [1] have shown that the final gene signature depends highly on the subset of patients used for the gene selection process. Later, the same team pointed out that the reproducibility of a signature depends on the number of samples used for the analysis [2]. Other teams were interested in the False Discovery Rate (FDR); that is, the expected proportion of false positives among the genes declared as significant. When looking for differential genes, Pawitan and al. showed that the FDR is mostly influenced by the proportion of truly differentially expressed genes and by the sample size [3]. Lee pointed out that to obtain reliable results there is a need to control simultaneously the power and type I error risk of the study [4]. The same problems are met in survival studies, where transcriptomic model construction raises simultaneously the problem of restricting the number of genes to be included and that of their validation. When a too complex model is fitted, - i.e it has too many free parameters to estimate for the amount of information in the data - the strength of the model will be exaggerated. This situation corresponds to overfitting. As a consequence, some conclusions of the analysis may be due to "noise" or to some spurious associations between the covariates and the outcome. This may lead to "optimism" that may be defined as overestimation of the ability of the model to predict outcome with new datasets. In this case, the model has the advantages of a training set; i.e., it has high adequacy and predictive accuracy but it is not able to predict efficiently the outcome of new datasets. The main objective of this article is to examine what happens when clinical variables and genes are introduced in the same model. The study was based on simulations in order to quantify to which extent the optimism of transcriptomic models may lead to overestimate the contribution of transcriptomic variables to prognosis, especially versus clinical variables.

2. METHODS

To compare optimism due to clinical and transcriptomic models within the context of survival, the study was based on simulated datasets that included both clinical and transcriptomic variables. The first step was the choice of the variables. Once chosen for clinical variables or selected for genes, these variables could be introduced in several models. To measure the predictive information contained in each survival model, we used the Kent and O'Quigley R^2 [5] measure of dependence. So, optimism for both types of variable could be compared.

2.1. SIMULATIONS

We considered a virtual population in which both clinical and transcriptomic variables were available for each patient. Two clinical variables were considered significant and were simulated using binomial distributions with probabilities 0.5 and 0.4, respectively, as parameters for success. Normal distributions $N(0,1)$ were assumed for transcriptomic variables. p genes were under study of whom only $p1$ genes were simulated to be truly associated with survival; the $p0$ remaining genes were simulated under the null hypothesis H_0 of no association with survival. Note that $p = p1 + p0$.

The $p+2$ variables were related to survival through a multivariate Cox model. In this model, coefficients are fixed at 0.8 for clinical variables, 0.2 for the $p1$ genes, and zero for the remaining $p0$ genes. A Weibull distribution with shape parameter 5 and scale parameter 2 was used for the baseline function. For censoring times, a uniform distribution $U(0,8)$ was used, leading to about 40% censoring.

For a fixed set of parameters p, p_1 60 training sets of n patients were simulated according to the design described above. For each of these training sets, 50 corresponding test sets were drawn following the same design. This overall process was performed varying sequentially n, p and p_1. The number of patients n was taken in {50,100,200,400}, p was considered in {500,1000,2000,4000} and p_1 was considered in {5,10,20}.

2.2. VARIABLE SELECTION AND MODEL CONSTRUCTION

Variable selection is used when no specific knowledge allows deciding on the most important predictors to include in the model among several potential predictors. As mentioned earlier, clinical variables were considered as validated and, thus, directly introduced in the models.
The classical way to link variables to censored survival data is to use the Proportional hazards Cox model. As mentioned earlier, clinical variables were considered as validated and, thus, directly introduced in such models. However, in the case of genes, the huge number of variables in comparison with the number of individuals prevents the maximization of the partial likelihood and the traditional Cox model can not be used directly. We used an extension to survival of the Threshold Gradient Descent method to overcome this problem ([7], [8]).
Let first recall briefly the Proportional hazards Cox model. Let us define X an (n, m) matrix of m variables for n individuals. For each of the n patients, the follow-up times are noted $t_1, ..., t_n$ and the event-indicators $d_1, ..., d_n$ with $d_i = 1$ if the event occurred and $d_i = 0$ if it did not occur. The Cox proportional hazards model is given by:

$$\lambda(t|X) = \lambda_0(t)exp(\beta' X) \tag{1}$$

where $\lambda_0(t)$ is a baseline hazard function, $\beta = \{\beta_1, ..., \beta_m\}$ is the vector of parameters and $X_1, ..., X_n$ are the vectors of gene expression levels for each of the p genes. In the Cox model, the vector of parameters is such that it maximizes the following the Cox partial likelihood (PL):

$$PL(\beta) = \prod_{k \in D} \frac{exp(\beta' \mathbf{x}_k)}{\sum_{j \in R_k} exp(\beta' \mathbf{x}_j)} \tag{2}$$

where D is the set of indices of the events and R_k is the set of indices of the individuals at risk at time t_k. Let us define $l(\beta) = -logPL(\beta)$.

Several methods have been proposed to deal as reliably as possible with survival models involving genes. Two main solutions are classically proposed within the transcriptomic context. The first solution reduce the space dimension using linear combinations of genes ([9], [10] or [11]). The second one shrink the estimated parameters, which is a desirable pattern considering the potential optimism in high dimension data ([12], [13], [14]). In the shrinkage approach, model selection lays on a path selection in the parameter space, this path being defined between the intercept as starting point and the full model as ending point. The Threshold Gradient Directed path method (TGD) directly constructs the path in the parameters space, in a sequential manner. Recently, [8] extended this gradient descent method proposed by Friedman and Popescu [7] to the Cox model. They demonstrated its ability to select relevant genes and the resulting predictive performance. We therefore decided to use this model to select the genes to introduce in our models.

With the TGD method, the path is constructed iteratively along the negative gradient of the partial log-likelihood defined as:

$$\mathbf{g}(\nu) = -\partial l / \partial \beta \tag{3}$$

The path begins with the intercept as starting point, that is, $\hat{\beta} = 0$. The parameter ν, which controls the number of iterations begins at zero too. The vector of parameters $\hat{\beta}$ is updated at each iteration:

$$\hat{\beta}(\nu + \Delta\nu) = \hat{\beta}(\nu) + \Delta\nu.h(\nu) \tag{4}$$

The steps are made in the direction of the gradient. $\Delta(\nu)$ controls the incremental moving along the gradient. $h(\nu)$ is defined as:

$$h(\nu) = \{f_j(\nu).g_j(\nu)\}_1^p \tag{5}$$

with

$$f_j(\nu) = I[|g_j(\nu)| \geq \tau.\max_{1 \leq k \leq p}|g_k(\nu)|] \tag{6}$$

$I[.]$ being the indicator function, and τ a user-defined constant $\in$ [0,1].

Through $\mathbf{f}_(\nu)$, only some coefficients are updated at each step, those for which the gradient exceeds a certain threshold determined by τ. For each step corresponding to a particular ν, the

cross-validated partial log-likelihood (CVPL) of the model is computed with the corresponding estimated $\hat{\beta}$. The value of ν that minimizes the CVPL gives the finally kept model. The corresponding vector of parameters $\hat{\beta}$ has only a few non-null coefficients that correspond to predictive genes. Finally, the TGD method combines selection of genes and estimation of their effect on survival. We used this method only for selection purposes; that is, to select genes irrespectively of their estimated coefficients.

Once chosen, clinical and transcriptomic variables are used in several models. The two clinical variables are kept in a $(n, 2)$ matrix X_C. Genes selected by the TGD are kept in an (n, k_T) matrix X_T, where k_T denotes the number of genes selected.

$$\lambda(t) = \lambda_{C0}(t) \exp(\alpha_C X_C) \tag{7}$$

$$\lambda(t) = \lambda_{T0}(t) \exp(\alpha_T X_T) \tag{8}$$

$$\lambda(t) = \lambda_0(t) \exp(\gamma_T X_T + \gamma_C X_C) \tag{9}$$

Clinical model (Equation 7) involves clinical variables in a multivariate Cox model. In the transcriptomic model, (Equation 8), coefficients of genes selected with the TGD are reestimated. The adjusted model (Equation 9) combines clinical and transcriptomic variables. Below, models 7 and 8 will be called the "clinical" and the "transcriptomic" model, respectively, and used to compare the optimism obtained with the clinical and the transcriptomic variables. To achieve this goal we had to choose a tool that measures the predictive quality of the models and the contribution of the variables to the prognostic, as described below.

2.3. COMPARISON OF THE CONTRIBUTION OF THE VARIABLES TO THE PROGNOSIS

Different criteria allow selection and comparison of models based on their predictive quality. Among these criteria, [15] showed the particular interest of the R^2 from Kent et O'Quigley [5], [6] which they denoted ρ^2. In linear regression models, the R^2 measures the part of the variance explained by a model. Kent and O'Quigley proposed an extension of this criterion to Cox Proportional Hazards model, based on the Kullback-Leibler Information that measures information gain brought by the variables in a model.

We used the ρ^2 to estimate model predictive ability and compare optimism of different models involving clinical and/or transcriptomic variables. Three ρ^2 were computed: ρ^2_{exp}, ρ^2_{train}, and ρ^2_{test}.

Expected ρ^2, ρ^2_{exp}, was computed using coefficients of variables defined in the simulations. This model includes clinical and transcriptomic variables, so ρ^2_{exp} corresponds to an adjusted. It estimates a measure of predictive accuracy forcing all candidate factors into the fit. ρ^2_{train} was computed using models (7), (8) or (9) on the training sets. Genes were selected on the same datasets. As to ρ^2_{test} we had to re estimate on the the test set coefficients of genes selected on the training set. It allows to evaluate simultaneously the effect of variable selection and coefficient estimation on new data sets. Using the coefficients derived from the training set to compute ρ^2 of the test set would suppose that the same variables are relevant to predict survival for training and test sets. It would have measured the amount of information these

genes bring, given that they have the same effect on the test set. By doing so, the selection step would have been denied. $\bar{\rho}^2_{test}$ is the mean of the ρ^2_{test} computed on the 50 test sets. Model 9 is used to compute adjusted ρ^2. In the case of genes, it gives the ρ^2 attributable to genes when clinical variables are involved in the model too.

Computing differences between various ρ^2 was a way to quantify optimism. Δ_{TrTe} (Equation 10) gives the differences between ρ^2_{train} and ρ^2_{test}. Comparing ρ^2 between the training and the test sets shows the error done by considering that the signature given by one dataset is the real signature and delivers the same information on other datasets. In other words, it gives the difference the predictive information anticipated on one dataset and the effective information on another. Δ_{TrExp} (Equation 11) and Δ_{TeExp} (Equation 12) compared respectively adjusted ρ^2_{train} and ρ^2_{test} with ρ^2_{exp}. In a given dataset, both measures give the difference between the effective predictive information and the detected one. Comparing the expected ρ^2 to the adjusted ρ^2_{train} showed how the predictive information thought to be contained in training sets moved away from the true one. This allowed evaluating the validation process. Comparing the expected ρ^2 to the adjusted ρ^2_{test}, we was able to evaluate the selection process. These Δ were computed for each combination of the three parameters $p, p1$ and n.

$$\Delta_{TrTe} = \frac{\sum_i^{60} \left(\rho^2_{train\,i} - \bar{\rho}^2_{test\,i}\right)}{60} \tag{10}$$

$$\Delta_{TrExp} = \frac{\sum_i^{60} \left(\rho^2_{train\,i} - \rho^2_{exp\,i}\right)}{60} \tag{11}$$

$$\Delta_{TeExp} = \frac{\sum_i^{60} \left(\bar{\rho}^2_{test\,i} - \rho^2_{exp\,i}\right)}{60} \tag{12}$$

3. RESULTS

3.1. INFLUENCE OF PARAMETER VARIATIONS

3.1.1. NUMBER OF PATIENTS Results are shown in an example with $p = 1000$ genes. Figures 1 shows the differences between ρ^2_{train} and ρ^2_{test} (Δ_{TrTe}) for the transcriptomic (on the left), and the clinical (on the right) models. Regarding genes, the difference decreases with increasing sample size, whatever the value of $p1$. These differences never reach zero. The confidence intervals are constant whatever the number of patients; as they are broad and overlapping, $p1$ cannot be considered as influential. Regarding clinical variables, Δ_{TrTe} varies around zero and does not depend on the sample size. The more the number of patients increases, the narrower the confidence intervals are.
The differences follow the same scheme when one computes Δ_{TrTe} with adjusted clinical or transcriptomic models (Results not shown).
These results show that transcriptomic and clinical variables have very different behaviours so they cannot be interpreted the same way. The predictive power of genes selected on one dataset is overestimated with regards to the predictive power they would have with other datasets. On the contrary, the predictive power for clinical variables is the same on both training and test sets. These results show that transcriptomic and clinical variables have very different

behaviours so they cannot be interpreted the same way.
Figure 2 shows the differences between expected values ρ^2_{exp} and ρ^2_{train} for the transcriptomic (on the left), and the clinical (on the right) models. Regarding genes, Δ_{TrExp} tends to zero when the number of patients increases; the more samples there are, the nearer to the expected ρ^2 is the adjusted ρ^2_{train}. There is an effect of $p1$: the smaller it is, the more distant are ρ^2_{exp} and ρ^2_{train}. This shows that with too small sample sizes, the high predictive information assumed to be given by the selected genes is far from the true information. The TGD selects more genes than the number of genes truly related to survival. These genes are false positive but they contribute to the computation of ρ^2_{train}. This gives the illusion that the genes have a great prediction power but in fact they are noise. The ρ^2 measure is sensitive to the number of variables involved in the model; as a consequence, ρ^2_{exp} increases with increasing $p1$, though non linearly. On the contrary, we observed that the number of selected genes does not practically vary with $p1$ and is of the order of 10. So, ρ^2_{train} is the all the more distant from ρ^2_{exp} that $p1$ is low because, in this case, the model computed on the training set involves more genes than the theoretical model used for the simulations. Regarding clinical variables, their predictive power is all the more underestimated that the number of patients is low. Moreover, this observation is amplified with high values of $p1$. This can be explained by the bias due to missing covariates when misspecifying a model [17]. When explanatory variables are omitted in a non-linear model, the effect of non-missing covariates is underestimated. This is typically what happens when selecting genes on the training set and when some relevant genes are not detected by the method. Because the TGD method selects practically the same number of genes whatever $p1$, the selection misses all the more genes that $p1$ is high. Note that, on the contrary, including non-relevant genes in the model does not bias the estimation of the other covariates.

We were also interested in differences between ρ^2_{exp} and adjusted ρ^2_{test}. It shows that by using the genes selected with the training set in the test set, the differences Δ_{TeExp} are negative: this means that the selected genes are not able to report the true information contained in the dataset. In other words, the genes selected on the previous dataset have no predictive power on other datasets because they are not the true ones.

When selecting genes, genes can either be really related to survival (genes under the alternative hypothesis $H1$) or not (genes under the null hypothesis $H0$). The former are true positives (TP), and the latter false positives (FP). To study the influence of the TP on optimism, we compared the evolution of the ρ^2 due to the true positives (TP) on one hand, and to all selected genes on the other hand. This was done with the training and the 3test sets. Figure 1 (on the left) shows that increasing n, ρ^2_{train} remains of the same order for all selected genes whereas the ρ^2 due to the TP increases. On the contrary, the right panel of 1 shows that ρ^2_{test} for all selected genes or TP evolve in the same way. In cases with 50 or 100 patients, there are no TP; ρ^2_{train} is also only due to noise; this cannot be seen when using only one dataset for a study and may lead to interpret wrongly this noise as information. Note that confidence intervals for 50 and 100 patients are missing. For this low number of patients, the lower limit of the confidence interval is null, because of the low number of TP.

3.1.2. TOTAL NUMBER OF GENES Results are shown in an example with $p = 100$ patients. Figure 4 shows the values of the differences between ρ^2_{train} and ρ^2_{test} in the transcriptomic model. When total number of genes increases, Δ_{TrTe} increases. When there are too few genes of interest, the selection method has difficulties to find the good ones. Genes

selected on the training set have no predictive power on the test sets. Given the overlapping of the confidence intervals, the effect of $p1$ can be considered as not influential.

The study of Δ_{TrExp} (results not show here) indicates that the more genes there are, the higher are the differences between ρ^2_{train} and ρ^2_{exp}, even more than $p1$ is high. It indicates that the more genes there are, the more optimism there is: : the predictive power of the transcriptomic model is all the more overestimated that the number of genes under study is high. The high value of ρ^2_{train} is due to noise and not real information. The high value of ρ^2_{train} is due to noise, and not real information. The study of Δ_{TeExp} (results not shown here), indicates that genes selected on the training set are not able to relay the predictive power really contained in the test set when the number of genes truly related to survival is too small relatively to the total number of genes. Indeed, differences are negative, even more than $p1$ is high.

4. DISCUSSION

4.1. ABOUT THE METHOD OF SELECTION OF GENES

To achieve our goal, we had to choose a gene selection method, the Threshold Gradient Descent method. Some comments may be made about this. First, from one dataset to another, the set of selected genes varies widely in number and identity. The number of TP that increases with the number of samples is more stable. One point that appears to us as a drawback of the TGD method is that it gives very low coefficient estimations, even for TP. Once genes were selected, coefficients of these genes were re-estimated in a new Cox model. We noticed that some of these genes had very low estimated coefficients. This may mean that the method is sometimes unable to really select genes, especially when the number of patients is low.

4.2. INFLUENCE OF PARAMETER VARIATIONS

By comparing the predictive powers of clinical variables and genes, two phenomena have to be taken into account: overestimation with genes due to the selection process and underestimation with clinical variables due to the omission of relevant genes. Genes are not yet validated. They are selected on a single dataset and assumed to have the same predictive power with other datasets. But the present results show that this predictive power is overestimated in the case of genes. This overestimation is all the more significant that the number of patients is low and the total number of genes is high. This is due to two overlapping phenomena: the gene selection mechanism and the power problem. When there are too few patients, the selected genes are not the true ones due to lack of power. This problem is not encountered with clinical variables that do not undergo a selection process because they have been already validated. The same problem arises when there are too many genes relatively to the number of genes truly related to survival. This problem is difficult to solve in genuine studies and must be kept in mind. Indeed, the number of genes of interest is not known in advance. This should be kept in mind when introducing clinical and transcriptomic variables in the same model. The predictive power of the clinical variables should not be neglected. In comparison with genes, their importance is not overestimated, which gives the feeling that they have less influence while their impact is hidden by the optimism encountered with genes.

5. CONCLUSION

With clinical variables, there is no need for selection of relevant variables; the optimism is low. On the contrary, with genes, the selection process introduces a high optimism. This optimism increases when the proportion of genes of interest decreases; that is, the total number of genes increases. That optimism can be decreased by increasing the number of samples. All these remarks show that the real effect of genes is overestimated compared to that of clinical variables.

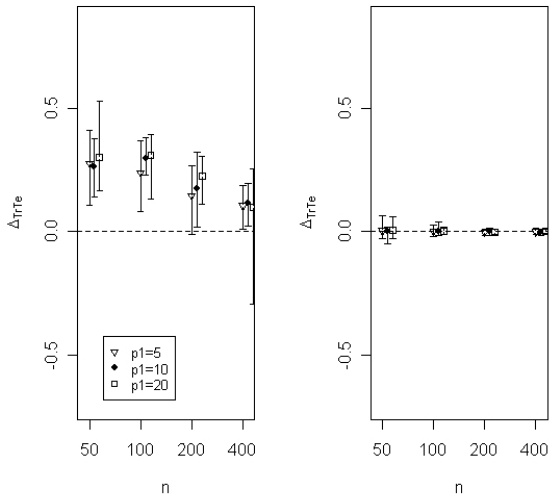


Figure 1. Evolution of Δ_{TrTe}, the difference between ρ^2_{train} and ρ^2_{train}, with n for the transcriptomic (on the left), and the clinical (on the right) models- $p = 1000$ genes. Confidence interval are represented by the vertical arrows. Each type of point symbol corresponds to a number $p1$ of genes truly related to survival.

ACKNOWLEDGEMENTS

We wish to thank Jean Iwaz for editing the manuscript.

REFERENCES

1. Ein-Dor,L. et al. (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171-178.
2. Ein-Dor,L. et al. (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences*, **103**, 5923-5928.
3. Pawitan,Y. et al. (2005) False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics*,**21**, 3017-3024.

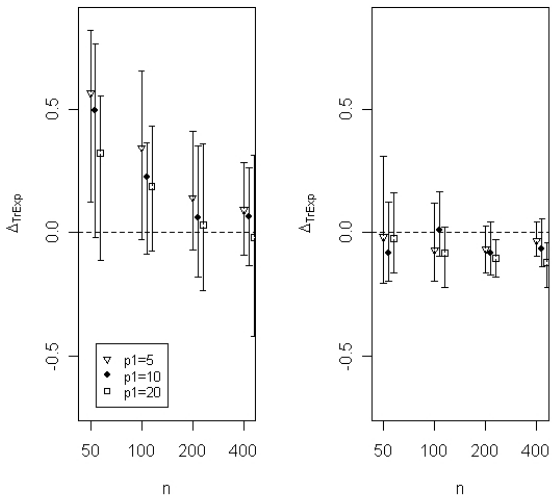


Figure 2. Evolution of Δ_{TrExp}, the difference between ρ^2_{train} and ρ^2_{exp}, with n for the transcriptomic (on the left), and the clinical (on the right) adjusted models - $p = 1000$ genes

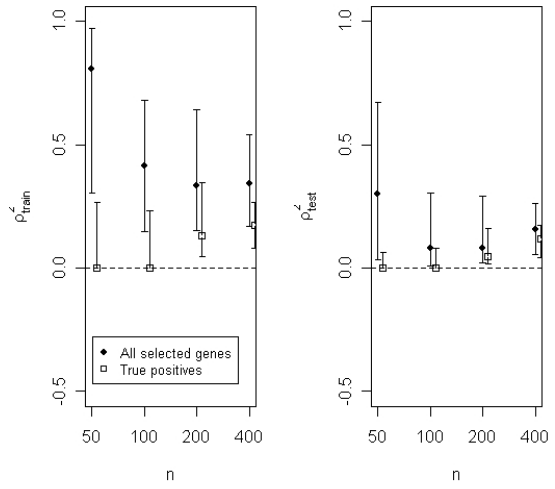


Figure 3. Evolution of ρ^2_{train} (on the left) or ρ^2_{test} (on the right) with n given that all selected genes or only TP are taken into account - $p = 1000$ genes

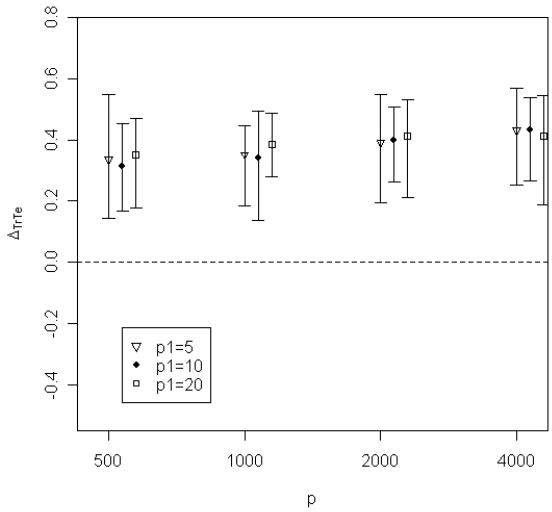


Figure 4. Evolution of Δ_{TrTe}, the difference between ρ^2_{train} and ρ^2_{test}, with p for the transcriptomic (on the left), and the clinical (on the right) models- $n = 100$ patients

4. Lee,M.LT, Whitmore, G.A (2002) Power and sample size for DNA microarray studies. *Statistics in Medicine*, **21**, 3543-3570.
5. Kent,J.T. and O'Quigley,J. (1988) Measures of dependence for censored survival data. *Biometrika*, **75**, 525-534.
6. O'Quigley J et al. (2005) Explained randomness in proportional hazards models. *Statistics in Medicine*,**24**, 479-489.
7. Friedman,J.H. and Popescu,B.E. (2004) Gradient directed regularization for linear regression and classification. *Technical Report, Department of Statistics, Stanford University*, **21**.
8. Gui,J. and Li,H. (2005) Threshold gradient descent in methods for censored data regression with applications in pharmacogenomics. *Pacific Symposium on Biocomputing*, 17-28.
9. Li L. and Li,H. (2004) Dimension reduction methods for microarrays with application to censored survival data. *Bioinformatics*, **20**, 3406-3412.
10. Bair,E. and Tibshirani,R. (2004) Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. *PLOS Biology*, **2** (4), 511-522.
11. Nguyen,D.V and Rocke,D.M. (2001) Assessing patient survival using microarray gene expression data via Partial Least Squares proportional hazard regression. *Computing Science and Statistics*,**33**, 376-390.
12. van Houwelingen, H.C et al. (2006) Cross-validated Cox regression on microarray gene expression data. *Statistics in Medicine*, **25**, 3201-3216.
13. Li, H. and Luan,Y. (2003) Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing*, **8**, 65-76.
14. Gui,J. and Li,H. (2005) Penalized Cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics*, **21**, 3001-3008.
15. Maucort-Boulch,D. et al (2006) Susceptibility to censorship of predictive accuracy measures. *submitted*.
16. Kullback,S. and Leibler,R.A (1951) On information and sufficiency. *Annals of Mathematical Statistics* , **22**, 79-86.
17. Chastang,C et al. (1988) GA quantitative study of the bias in estimating the treatment effect caused by omitting a balanced covariate in survival models. *Statistics in Medicine*, **7**, 1243-1255.

Annexe C

Glossaire

Akaike's Information Criterion (AIC) : Mesure de la qualité d'ajustement d'un modèle en permettant un compromis entre la complexité du modèle et son ajustement aux données. La mesure repose sur une estimation de la distance entre le meilleur modèle, supposé existent, et le modèle testé.

Démembrement des lymphomes à grandes cellules B (DLBCL) : un des sous-types les plus fréquents des lymphomes malins, représentant 30 à 40 % des lymphomes de l'adulte (Diffuse Large B Cell Lymphoma en anglais).

Hybridation : Association de chaînes d'acides nucléiques simple brin pour former des doubles brins, basée sur la complémentarité des séquences de nucléotides.

IPI (International Prognostic Index) : Outil clinique développé par les oncologues pour aider à prédire le pronostic de patients atteints d'un lymphome non Hodgkinien agressif. Cet index intègre les critères suivants : l'âge (>60 ans ou non), le statut de la maladie, le nombre de sites extra-nodaux, le taux de LDH, le niveau d'état de santé général.

Leucémie aiguë lymphoblastique : Leucémie aiguë caractérisée par la prolifération incontrôlée de lymphocytes immatures dans le sang et la moëlle.

Leucémie aiguë myéloblastique : La leucémie aiguë myéloblastique est causée par un surnombre d'un autre type de cellules, les ńgranulocytesż. Ces cellules demeurent immatures et affectent les globules blancs, les globules rouges et les plaquettes de la même manière que les lymphocytes dans les cas de leucémies aiguës lymphoblastiques.

Maladie de Hodgkin : Affection cancéreuse caractérisée par une prolifération cellulaire anormale dans un ou plusieurs ganglions lymphatiques.

Myélome : Cancer hématologique de la moelle osseuse. C'est une maladie caractérisée par le développement dans toutes les parties du squelette de multiples tumeurs ostéolytiques.

Oligonucléotide : Segment d'ADN simple brin composé de quelques dizaines de nucléotides.

Transcription inverse : Ensemble des mécanismes moléculaires conduisant à la synthèse d'un ADN (Acide Désoxyribonucléique) à partir de l'ARN (Acide Ribonucléique).

Translocation $t(9;22)$: La translocation t(9 ;22) correspond à un échange de segments entre les gènes ABL du chromosome 9 et BCR du chromosome 22. Cette translocation caractérise un sous-groupe des leucémies aiguës lymphoblastiques : les leucémies myéloïdes chroniques.