Robustess of Phylogenetic Trees

Mahendra Mariadassou

N° d'ordre: 9658

THÈSE

Présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS-SUD XI

Spécialité: Mathématiques

par

Mahendra MARIADASSOU

Robustesse des arbres phylogénétiques

Soutenue le 27 Novembre 2009 devant la Commission d'examen:

M.	Avner BAR-HEN	(Directeur de thèse)
M.	Philippe BESSE	(Rapporteur)
M.	Olivier GASCUEL	(Rapporteur)
M.	Hirohisa KISHINO	(Co-encadrant)
M.	Pascal MASSART	(Examinateur)
Mme.	Patricia REYNAUD-BOURET	(Examinateur)

Thèse préparée au
Département de Mathématiques d'Orsay
Laboratoire de Mathématiques (UMR 8628), Bât. 425
Université Paris-Sud 11
91 405 Orsay CEDEX

Robustness of Phylogenetic Trees

Abstract

Phylogenetic trees are used daily in many fields of biology, most notably the functional and structural study of genomes. They provide a powerful framework to study evolution but are also an abundant source of statistically challenging issues. Most, if not all, applications of phylogenetics have in common that they require accurate phylogenetic estimates. In general, accurate estimates depend on four factors: (1) appropriate selection of genes, (2) sufficient data size, (3) accurate analytical method, (4) adequate taxon sampling. We present in this thesis four issues directly related to this factors. In the first part, we use concentration inequalities to upper bound the amount of data needed to choose the most accurate of two trees when the analytical model is accurate. Using Edegworth expansions, we then present a procedure to select congruent genes from a list of target genes. In the second part, we propose two procedures, based on influence function and sensitivity curves, to identify influent nucleotides and taxa, which are likely to impede the inference and lead to non-robust estimates. We show that as few as one nucleotide or taxon can have a drastic impact on the estimates, discuss the biological implication of this result and provide methods to achieve greater robustness of the trees.

Keywords : Phylogenetics, Robustness, Edgeworth expansions, Concentration inequalities, Influence function, Outliers detection, Taxon Influence Index.

Résumé

La théorie synthétique de l'évolution a largement diffusé dans tous les domaines de la biologie, notamment grâce aux arbres phylogénétiques. S'ils ont une utilité évidente en génomique comparative, ils n'en sont pas moins utilisés dans de nombreux autres domaines allant de l'étude de la biodiversité à l'épidémiologie en passant par les sciences forensiques. Les arbres phylogénétiques sont non seulement une charactérisation efficace mais aussi un outil puissant pour étudier l'évolution. Cependant, toute utilisation d'arbre dans une étude suppose que l'arbre ait été correctement estimé, tant au niveau de la topologie que des autres paramètres, alors que cette estimation est un problème statistique compliqué et encore très ouvert. On admet généralement qu'on ne peut faire de bonne estimation sans les quatre pré-requis que sont (1) le choix d'un ou plusieurs gènes pertinents pour la question étudiée, (2) une quantité suffisante de données pour s'assurer une bonne précision d'estimation, (3) une méthode de reconstruction efficace qui s'appuie sur une modélisation fine de l'évolution pour minimiser les biais de reconstruction, (4) un bon échantillonnage de taxons. Nous nous intéressons dans cette thèse à quatre thèmes étroitement liés à l'un ou l'autre de ces pré-requis. Dans la première partie, nous utilisons des inégalités de concentration pour étudier le lien entre précision d'estimation et quantité de données. Nous proposons ensuite une méthode basée sur des extensions de Edgeworth pour tester la congruence phylogénétique d'un nouveau gène avec ses prédécesseurs. Dans la deuxième partie, nous proposons deux méthodes, inspirées des analyses de sensibilités, pour détecter les sites et taxons aberrants. Ces points aberrants peuvent nuire à la robustesse des estimateurs et nous montrons sur des exemples comment quelques observations aberrantes seulement suffisent à drastiquement modifier les estimateurs. Nous discutons les implications de ces résultats et montrons comment augmenter la robustesse de l'estimateur de l'arbre en présence d'observations aberrantes.

Mots-clefs : Arbres phylogénétiques, robustesse, développement de Edgeworth, inégalités de concentration, détection de points aberrants.

Remerciements

Comme dit l'adage, "À tout seigneur, tout honneur", ces remerciements ne pourraient commencer avec personne d'autre qu'Avner Bar-Hen. Je le remercie pour ces quelques années passées à le côtoyer et à travailler sous sa direction. Sur le plan professionnel, je le remercie de m'avoir amené au domaine de la phylogénie. Il m'a permis par sa disponibilité, sa capacité d'écoute et ses conseils de mener à bien cette grande aventure humaine qu'est la thèse. Ses encouragements et son enthousiasme permanent ont été les meilleurs remèdes contre les doutes et les phases de découragement scientifique qui m'ont assailli en cours de thèse. Sur le plan personnel, j'ai beaucoup profité et appris de quatre années de discussions sur des sujets sérieux et d'autres plus légers. Je le remercie chaleureusement pour son coté facétieux et sa curiosité, qui m'ont permis développer la mienne. Enfin, je le remercie pour les nombreuses fois où il m'a fait profiter de sa connaissance des bonnes tables parisiennes. Puissent ses futurs thésards en profiter autant que moi !

Je remercie également chaleureusement Hirohisa Kishino pour son accueil fantastique. Non content de me fournir un cadre de travail propice à la réflexion, il m'a appris l'autre moitié de ce que je sais en phylogénie. Ses références justes et sa connaissance encyclopédique du domaine et des gens qui y travaillent se sont avérées précieuses à de nombreuses reprises. Ses méthodes originales, comme le "café des sciences" matinale, et sa gentillesse m'ont permis de travailler dans un environnement convivial et stimulant. Enfin, je le remercie tout particulièrement d'avoir fait 24 heures d'avion en trois jours pour participer à mon jury de thèse.

Je voudrais remercier les membres du jury. Tout d'abord Philippe Besse et Olivier Gascuel pour m'avoir fait le grand honneur de rapporter ma thèse. Ils ont passé de (longues) heures à lire le manuscrit et, via leurs commentaires, m'ont proposé de nouvelles idées et des perspectives très intéressantes.

Je remercie également Patricia Bouret. Du mémoire de maîtrise, où elle m'a fait découvrir les tests, à la fin de la thèse où elle m'a aidé à améliorer le manuscrit, elle a fait preuve à mon égard d'une constante bienveillance.

Pascal Massart a accepté de présider mon jury ; qu'il en soit remercié. Ses remarques ont donné lieu à une discussion scientifique très riche lors de la soutenance.

Je voudrais également remercier les chercheurs confirmés ou en formation qui m'ont aidé dans mes recherches, en m'aiguillant, en me fournissant des références ou en initiant une conversation intéressante : Marie-Anne Poursat, Phlippe Vandenkoornhuyse, Leonardo de Oliveira Martins, Ziheng Yang, Jeff Thorne, Elcio Leal. Je remercie également, dans un registre moins lié à la phylogénie mais tout aussi chaleureusement, Stéphane Robin et Jean-Jacques Daudin. Je remercie aussi les collègues qui m'ont encouragé durant la thèse par leur soutien et leur bonne humeur permanente, ou en partageant leur expérience (pas toujours lointaine) de thésard. Je pense à Tristan Mary-Huar, Emilie Lebarbier, Michel Koskas, Marie-Laure Martin-Magniette, Liliane Bel, Colette Vuillet, Tae-Kun Seo, Servane Gey et Adeline Samson.

J'ai eu la chance d'effectuer ma thèse sur trois laboratoires différents. Cette situation atypique n'a pas été sans poser son lot de problèmes administratifs mais j'ai bénéficié de très bonnes conditions matérielles partout et de l'aide de nombreuses personnes pour résoudre les problèmes : Odile Jalabert et Carole Tiphaine à l'Agro, Mieko Sasaki à Tokyo, Michel Guillemot et Marie-Hélène Gbaguidi au MAP5. Je ne peux qu'être

reconnaissant à la clique des thésards, à mi-chemin entre amis et collègues, avec qui j'ai partagé un bureau ou un labo. Je pense évidemment à Alain, son amour du Nord, ses chemises manches courtes par -2 et nos conflits sur la "bonne" température du bureau. Merci à Shun et Leo pour les petits conseils sur les trucs à faire et la vie à Tokyo. Merci aux (nombreux) thésards et assimilés du MAP5 : Cécile, Benjamin, Cheikh, Anne-Cécile, Sandra, Emeline, Nathalie K., Nathalie A., Jean-Patrick, Arno, Georges, Solange, Jean-Pascal et Baptiste (pas forcément dans cet ordre) pour m'avoir fait découvrir le sens profond des mots gentillesse, curiosité, enthousiasme, bonne humeur, ouverture d'esprit, humour, franche rigolade, second degré, ironie, musique classique et cuisine épicée (pas forcément dans cet ordre). Tous ont contribué à rendre la thèse nettement plus sympa.

Ensuite, et parce que le labo n'occupe qu'une partie de mes journées, merci à tous les autres, amis et famille, qui m'ont soutenu pendant trois ans. Du côté français, merci à Matthieu et Chloé S. de me supporter depuis les Pays-Bas (10 ans déjà). Merci à Régis, Nico et Seb pour les soirées jeux/pizzas/bières/Wii. Merci à Chloé D. pour ses innombrables invitations à prendre un thé, les sorties gourmandes et les repas à base de fromage fondu. Merci à Guillem de faire la vaisselle à la fin des précédents repas. Merci à Marguerite et Liem Binh d'être toujours partant pour sortir, mais jamais libres en même temps. Merci à Pascal et Blanche pour les nombreux "Viens chez nous, on a un poulet/des restes à finir". Merci à Medhi et Yuko de partager mes passions un peu geeks. Merci à Sabine pour ses petites histoires sur le monde terrible des médecins. Merci à Nathalie, Merlin, Matthieu L. et Vincent de supporter mes blagues depuis le M2. Merci à Natacha, Ariane, Elifsu et Etienne de supporter mes blagues (les mêmes) depuis le M2 (l'autre). Du côté japonais, merci à Celulas et à tous ses membres pour leur soutien ininterrompu durant mon séjour au Japon, tout particulièrement Kachusha, Dori, Coppelia, Hachi, Hiro. Merci à Maiko, Ecchan, Naoko et les membres de AIKOM 13 our les sorties dans le Tokyo jeune. Merci à toute la famille Kurakata pour les sorties dans le Tokyo un peu moins jeune. Le tableau serait incomplet sans une mention spéciale à Kaoru et Don Ho pour les okonomiyakis et une mention toute spéciale à Yong Eun pour tout le reste. Un peu à cheval entre France et Japon, merci à Baptiste de garder le contact malgré la distance et d'être le contradicteur le plus retors du monde. Merci aussi à Eiichiro Oda et Luffy pour la dose hebdomadaire de bonne humeur.

Un très gros merci plus que mérité à toute ma famille, qui s'est littéralement préoccupée de mon bien-être à travers les continents. Même en période irascible de rédaction, leur soutien a été indéfectible. Mention spéciale à mon papa et ma maman, sans lesquels je serai mort de faim pendant la rédaction et mon oratoire serait mort de faim pendant le pot de thèse.

Enfin, j'embrasse ma princesse, celle dont la simple présence illumine ma vie et rend chaque jour un peu plus beau.

Contents

1 Introduction **14**

1.1 Le contexte de la phylogénie moléculaire 14

1.1.1 Origines du domaine 14

1.1.2 Méthodes de reconstruction d'arbres phylogénétiques 15

1.1.3 Validation de l'arbre 16

1.2 La vraisemblance en phylogénie 17

1.2.1 Modèle d'évolution 17

1.2.2 Calcul de la vraisemblance 23

1.2.3 Maximisation de la vraisemblance 24

1.3 Recherche de l'arbre 25

1.4 Validation de l'arbre phylogénétique 26

1.4.1 Test d'hypothèse de paramètre continu 27

1.4.2 Test de topologies 27

1.5 Plan de thèse 28

1.5.1 Variabilité normale du modèle 29

1.5.2 Détection des données aberrantes 30

General Introduction **14**

2 Introduction to Phylogenetics **31**

2.1 A bit of Context 31

2.1.1 A Brief History of Molecular Phylogenetics 31

2.1.2 Molecular Phylogenetics in Biology 33

2.1.3 Methods for Inferring Trees 34

2.1.4 Validating the tree 35

2.2 Likelihood function in Molecular Phylogenetics . . . 36
2.2.1 Three parts of an evolutionary model . . . 36
2.2.2 Markov models of sequence evolution . . . 38
2.2.3 Computing the Likelihood . . . 42
2.2.4 Optimizing the Likelihood . . . 47
2.3 Tree Search Strategies . . . 48
2.3.1 Complete Search . . . 48
2.3.2 Heuristic Searches: Initial Tree . . . 49
2.3.3 Heuristic Searches: Hill Climbing . . . 49
2.4 Validation and Significance of the Result . . . 52
2.4.1 Phylogenetic Model . . . 52
2.4.2 Hypothesis Testing . . . 53
2.4.3 Testing Topologies . . . 54
2.5 Thesis Outline . . . 55
2.5.1 Stochastic Errors . . . 55
2.5.2 Detecting Outliers . . . 56

I Stochastic Errors **59**

3 Introduction **61**
3.1 Concentration Inequalities for Sums of Independent Random Variables 61
3.2 Evolutionary Trees . . . 64
3.3 Consistency of the Generating Process Along the Sequence . . . 65

A Concentration Inequality for Evolutionary Trees **67**
A.1 Introduction . . . 68
A.2 Framework . . . 69
A.2.1 Notations and definitions . . . 69
A.2.2 Connection between ℓ^m and Q . . . 71
A.2.3 Distance between Q and Q_n . . . 73
A.3 Phylogenetic reconstruction for finite size samples . . . 74
A.3.1 Distance between the empirical and true mean log-likelihoods . 74
A.3.2 Support given to a tree . . . 76

A.4 Comparison with widely used methods and illustration 78
A.4.1 Bootstrap 78
A.4.2 Illustration of the method on an example 80
Appendix: Proof of the lemmas 83

B Assessing the Distribution Consistency of Sequential Data **84**
B.1 Introduction 85
B.2 Definitions and Notations 87
B.2.1 Definition of $\Delta_{n,+k}$ and $\Delta_{n,-k}$ 87
B.2.2 Characteristic Function 88
B.3 Edgeworth Expansion 89
B.3.1 With Cramér's Condition 89
B.3.2 Without Cramér's Condition 90
B.3.3 New Generating Process 90
B.3.4 About Discrete Distributions 91
B.4 Application to Test 92
B.4.1 Distribution of $\Delta_{n,+k}$ under H_0 93
B.4.2 Distribution of $\Delta_{n,+k}$ under H_1 93
B.4.3 Calibrating the test 94
B.4.4 Discussion of the results 94
B.5 Proofs 96
B.5.1 Previous Results 96
B.5.2 New Results 97
B.5.3 Proof of Prop 18 100
B.5.4 Proof of Theorem 15 101
B.5.5 Proof of Theorem 16 102
B.5.6 Proof of Theorem 17 103

4 Discussion and Prospects **105**
4.1 Summary 105
4.1.1 Concentration Inequalities and Gaussian Approximation 105
4.1.2 The Pitfall of Increasing Sequence Length 107
4.2 Further Work 108

II Detecting Outliers **116**

5 Introduction **118**

5.1 The hardships of confidence study in phylogenetics 118

5.2 Robustness Analysis 120

5.3 Towards Specific Robustness Indicators 123

C Influence Function **127**

C.1 Introduction 128

C.2 Methods 129

C.3 The dataset 132

C.4 Results and Discussion 132

C.5 Acknowledgments 136

D Taxon Influence **137**

D.1 Introduction 138

D.2 Material and Methods 139

D.2.1 Methods 139

D.2.2 Material 140

D.2.3 Phylogenetic Analysis 141

D.3 Results 141

D.3.1 Inference Quality 141

D.3.2 TII Distribution and Influential Taxa 142

D.3.3 Branch Stability 143

D.4 Discussion 143

D.4.1 Influential Taxa and Rogue Taxa 143

D.4.2 TII and BS scores 145

D.4.3 TII and Long Branches 145

D.4.4 Relation with bootstrap support 146

D.4.5 Extension to Bayesian methods 146

D.4.6 Limitations and future work 147

D.5 Acknowledgements 148

6 Discussion and Prospects **149**

6.1 Summary 149
6.2 Further Work 152

General Conclusion **159**

7 Conclusion **159**
7.1 Summary 159
7.2 Perspectives 160

Bibliography **163**

General Introduction

Chapitre 1

Introduction

Ce chapitre constitue une introduction au domaine de la phylogénie moléculaire et présente le sujet de thèse : la robustesse des arbres phylogénétiques. Nous commençons par une brève présentation historique du domaine avant de passer en revue les méthodes de reconstruction les plus populaires. Nous nous intéressons tout particulièrement à la méthode du maximum de vraisemblance. Cette méthode nécessite de construire un modèle probabiliste d'évolution des macromolécules biologiques mais fournit en contrepartie un cadre statistique propice à quantifier la variabilité de l'arbre estimé. Nous présentons tout d'abord les modèles d'évolution couramment utilisé, puis le calcul de la vraisemblance avant de montrer que la nature discrète de l'arbre rend caducs les outils traditionnels d'étude de la variabilité. Nous faisons un état de l'art des tests d'arbres basés sur la vraisemblance avant de présenter les résultats de la thèse.

1.1 Le contexte de la phylogénie moléculaire

1.1.1 Origines du domaine

Les travaux précurseurs de Charles Darwin Darwin (1859), sur lesquels a été bâtie la biologie évolutive moderne, ont radicalement changé notre compréhension de l'évolution. Darwin introduit dans son livre *De l'Origine des Espèces* la théorie de l'évolution, selon laquelle les espèces évoluent au fil des générations grâce au processus de sélection naturelle et que la diversité du vivant est obtenue grâce à l'accumulation graduelle de différences dans les sous-populations d'une espèce.

L'évolution peut être considérée comme un processus de branchement dans lequel des sous-populations d'une espèce se transforment par accumulation de différences avant de se détacher de leur espèce-mère pour former une nouvelle espèce ou s'éteindre. L'image d'arbre évolutif illustre bien le concept d'évolution et la formation de nouvelles espèces à partir d'espèces déjà existantes. Les liens de parenté qui unissent un groupe d'espèces sont communément représentés sous la forme d'arbres évolutifs, appelés "arbres phylogénétiques" ou encore "phylogénies".

Toutes les méthodes de reconstruction d'arbres phylogénétiques sont basées sur

la même idée intuitive : étant donné que l'évolution intervient par accumulation de différences, deux espèces qui ont divergé récemment sont plus "semblables" que deux espèces dont la divergence est plus ancienne. La similitude entre espèces était mesurée par des critères de types morphologiques (à l'instar de la forme des os, du nombre de pattes ou du nombre de dents) jusque dans les années 50. La découverte de la structure de l'ADN par Watson et Crick en 1953 (Watson and Crick, 1953) et surtout les capacités de séquençage et d'analyse des molécules macrobiologiques qui ont rapidement suivis ont considérablement changé la donne en remplaçant avantageusement l'objet d'étude. Au lieu d'établir des liens de parenté à partir de critères morphologiques, pour certains fortement soumis à l'appréciation de l'expérimentateur et dont le nombre est généralement faible, les phylogénéticiens peuvent désormais s'appuyer sur des données moléculaires : des séquences génétiques (d'ADN) ou protéiques (de protéines). Cette révolution présente trois avantages majeurs. Tout d'abord, l'évolution agit beaucoup plus finement au niveau moléculaire qu'au niveau des caractères morphologiques : certaines mutations de la séquence d'ADN sont invisibles au niveau morphologique. Ensuite, les séquences moléculaires sont moins soumises à la subjectivité de l'expérimentateur que les critères morphologiques. Enfin, les séquences moléculaires fournissent des jeux de données bien plus important que les critères morphologiques : au lieu de comparer les espèces sur quelques dizaines de critères morphologiques, on les compare sur des séquences longues de plusieurs milliers de paires de bases, voire de plusieurs millions pour les espèces dont l'intégralité du génome est connue.

Reconstruire l'histoire évolutive des espèces constitue évidemment un but en soi pour les biologistes évolutifs. Le symbole le plus emblématique en est le projet "Arbre de la Vie" (Tree of Life Project, `www.tolweb.com`), qui cherche à reconstruire l'arbre phylogénétique de toutes les espèces vivantes. Mais les arbres phylogénétiques ont aussi un intérêt majeur dans d'autres domaines de la biologie. Ils sont par exemple inestimables en génomique comparative, où ils permettent par exemple de prédire la fonction d'un gène inconnu à partir de la fonction d'un gène similaire dans des espèces proches (Eisen, 1998; Eisen and Wu, 2002) ou encore de prédire si deux protéines interagissent à partir de leurs arbres phylogénétiques respectifs (Pazos and Valencia, 2001). Mais le domaine d'application de la phylogénie ne se réduit pas à la biologie moléculaire : les phylogénies apparaissent aussi naturellement en biologie de la conservation quand, en particulier dans les études de mesure de la biodiversité (Bordewich et al., 2008).

1.1.2 Méthodes de reconstruction d'arbres phylogénétiques

Toutes les applications décrites dans la section 1.1.1 s'appuient sur des arbres phylogénétiques bien reconstruits. Mais reconstruire de tels arbres est une tache ardue : il s'agit de reconstituer le chemin parcouru par l'évolution à partir des empreintes qu'elle laisse sur les génomes, en sachant que ces empreintes peuvent être ténues et s'atténuent de toutes façons au fil du temps. Les systématiciens n'en reconstruisent pas moins des arbres évolutifs depuis Darwin, avec une précision étonnante.

Il existe essentiellement 5 grandes familles de méthodes pour reconstruire une phylogénie : les méthodes de parcimonie (Edwards and Cavalli-Sforza, 1963), les

méthodes de moindres-carrés (Cavalli-Sforza and Edwards, 1967), les méthodes de maximum de vraisemblance (Edwards and Cavalli-Sforza, 1964; Felsenstein, 1973), les méthodes de distance (Fitch and Margoliash, 1967) et les méthodes bayésiennes (Li, 1996; Mau, 1996; Rannala and Yang, 1996). La contribution majeure des travaux de Cavalli-Sforza et Edwards, tous deux disciples de Fisher, est sans doute l'identification précoce de la reconstruction d'arbres phylogénétiques comme un problème d'inférence statistique.

Toutes les méthodes évoquées ci-dessus peuvent être décomposées en trois parties :

1. Un *critère d'optimalité*, qui mesure l'adéquation des données à un arbre phylogénétique donné (par exemple : la parcimonie, la vraisemblance, les sommes de carrés, etc) ;
2. Une *stratégie de recherche* pour identifier l'arbre optimal (par exemple : recherche exhaustive, descente de gradient, etc)
3. Des hypothèses sur le *mécanisme d'évolution* des données.

Il n'existe pas de méthode supérieure à toutes les autres, chacune à ses forces et ses faiblesses et le débat sur les mérites comparés de deux méthodes n'est pas clos. Pour certains groupes d'espèces, le choix de la méthode importe peu : toutes les méthodes reconstruisent le même arbre phylogénétique. Il s'agit évidemment du cas optimiste, rarement rencontré en pratique. Nous nous intéressons dans cette thèse exclusivement à la méthode du maximum de vraisemblance. Cette méthode est nettement plus lente que ses concurrentes mais fournit un cadre de travail naturel tant pour tester des hypothèses que pour quantifier la variabilité de l'arbre estimé.

1.1.3 Validation de l'arbre

Comme dans la majorité des procédures d'inférence statistique, l'arbre estimé dépend des données : la même procédure d'estimation appliquée à différents jeux de données va donner différents arbres. Il est essentiel de quantifier cette variabilité, en prouvant par exemple que l'arbre estimé n'est pas très différent du vrai arbre. La façon standard de faire en est de prouver un théorème limite sur l'arbre estimé, généralement un théorème de normalité asymptotique sur la distance entre l'arbre estimé et le vrai arbre.

Mais un arbre phylogénétique est un paramètre inhabituel : il est composé d'une topologie (une forme d'arbre) discrète et de longueurs de branches continues, qui dépendent de la topologie de l'arbre. L'espaces des arbres phylogénétiques a de plus une structure complexe (Billera et al., 2001) qui rend inopérants les outils de convergence utilisés pour établir des théorèmes limites.

Faute de théorèmes limite, la variabilité de l'estimateur est généralement quantifiée à l'aide de technique de rééchantillonnages, telles que le bootstrap (voir section 2.4 et chapitre 5) ou le jackknife (voir chapitres 5 et C) qui miment des échantillons indépendants.

Enfin, il est nécessaire de valider la robustesse de l'arbre estimé. Les erreurs d'alignement et de séquençage peuvent en effet engendrer de petites modifications du jeu de données. Quelle est l'influence de ces petites modifications sur l'arbre estimé ?

Si leur influence est faible, l'arbre estimé est robuste aux erreurs de séquençage et d'alignement : il est légitime de s'en servir dans des analyses ultérieurs. Dans le cas contraire, l'arbre est peu robuste : les analyses basées sur cette arbre sont peu fiables. Là encore, les méthodes de bootstrap et de jackknife permette de quantifier la robustesse de l'arbre. Dans le cas d'arbres non robustes, il est intéressant d'identifier les données erronées pour les corriger ou les supprimer du jeu de données. Les erreurs de séquençage et d'alignements ont tendance à créer des données exceptionnelles, très différentes du reste du jeu de données et inattendu. Le cadre du maximum de vraisemblance permet non seulement de quantifier la variabilité de l'arbre estimé mais aussi le caractère exceptionnel ou non d'une donnée. Il est donc particulièrement propice à la détection de données erronées.

1.2 La vraisemblance en phylogénie

Nous nous intéressons à la vraisemblance comme critère d'optimalité. Calculer une vraisemblance nécessite de définir un modèle probabiliste du phénomène d'intérêt, ici l'évolution des séquences moléculaires. Nous présentons tous d'abord le modèle d'évolution avant de montrer que la vraisemblance est bien calculable et comment elle se calcule. Les modèles d'évolution pour les séquences protéiques étant très proches de ceux utilisés pour les séquences d'ADN, nous nous contentons des modèles d'évolution pour les séquences d'ADN.

1.2.1 Modèle d'évolution

Les modèles d'évolution pour les séquences d'ADN sont paramétrés par :

1. Un arbre phylogénétique ;
2. Un mécanisme d'évolution des données (généralement un processus Markovien d'évolution des nucléotides).

Commençons par définir un arbre phylogénétique.

Définition d'un arbre phylogénétique On appelle graphe $G = (V, E)$ est un ensemble V de noeuds et $E \subset V \times V$ de branches. Une branche $e = (v_1, v_2)$ est *incidente* à v_1 et v_2. Deux branches sont *adjacentes* si elles sont incidentes à un même noeud. Un *chemin* c est une suite de branches $e_1, \ldots, e_n$ telles que e_i est adjacent à e_{i+1} pour tout $i \leq n-1$. Si $e_1 = (v_0, v_1)$ et $e_n = (v_{n-1}, v_n)$, v_0 et v_n sont les *extrémités* du chemin et c relie v_0 et v_n. G est *connecté* si pour toute paire de noeuds (v_1, v_2), il existe un chemin reliant v_1 à v_2. Un chemin est *cyclique* si ses deux extrémités sont confondues. G est *acyclique* si il n'existe aucun chemin cyclique. Le *degré* d'un noeud est le nombre de branches incidentes à ce noeud. Les noeuds de degré 1 sont des *feuilles*, les autres sont des *noeuds internes*.

Un arbre binaire est un graphe acyclique connecté dont tous les noeuds internes sont de degrés 3, sauf un qui est de degré 2. Le noeud de degré 2 est la *racine*. Les branches incidentes à une feuille sont des *branches terminales*, les autres sont des

Baleine Finnoise	M	N	E	**N**	L	F	A	P	F
Baleine Bleue	M	N	E	**N**	L	F	A	P	F
Chimpanzé	M	N	E	**N**	L	F	A	S	F
Bonobo	M	N	E	**N**	L	F	A	S	F
Gorille	M	N	E	**N**	L	F	A	S	F
Orang-outan	M	N	E	**D**	L	F	T	P	F

TABLE 1.1 – : Exemple d'alignement de caractères de taille $s \times n$ (6 espèces $\times$ 10 caractères). X_4 est mise en avant.

branches internes. Pour toute paire de noeuds (v_1, v_2), il existe un unique chemin $c(v_1, v_2)$ reliant v_1 à v_2.

Un **arbre phylogénétique** est composé de :
- Une topologie T : un arbre binaire $T = (V, E)$ dont les feuilles sont étiquetées de $S_1, \ldots, S_n$. Chaque feuille représente une espèce.
- Des longueurs de branches $(t_e)_{e \in E} \leq 0$: t_e est la longueur de la branche e. On impose $t_e > 0$ pour les branches internes mais autorise $t_e = 0$ pour les branches terminales à condition que deux branches terminales adjacentes ne soient pas de longueur nulle en même temps.

Avant de présenter le mécanisme d'évolution des données, spécifions la forme des données.

Matrice de caractères En phylogénie, les observations sont une matrice de caractère alignés (ou alignement) de taille $s \times n$ où s est le nombre d'espèces et n est le nombre de caractères observés, comme illustré en Table 1.1. L'alignement est noté $\mathbf{X}$. Chaque ligne correspond à une séquence moléculaire est représente l'information dont on dispose pour une espèce. Chaque colonne de l'alignement représente un caractère (ou position, nucléotide ou site). Les observations individuelles sont les colonnes de la matrice, notées $X_1, \ldots, X_n$ de sorte que $\mathbf{X} = (X_1, \ldots, X_n)$. Dans notre exemple $X_4 = (N, N, N, N, N, D)'$: le caractère considéré vaut N chez toutes les espèces sauf l'orang-outan.

Construire un alignement n'est pas une tache aisée et l'alignement multiple de séquences constitue un domaine de recherche très actif en bioinformatique. Il est en particulier connu que les alignements de séquences dépendent fortement de la méthode d'alignement utilisé et que reconstruction phylogénétique et alignement de séquences sont étroitement imbriqués et devrait dans l'idéal être réalisés en même temps Liu et al. (2009). Nous laissons de coté le problème d'alignement dans cette thèse et considérons l'alignement comme donné.

Pour calculer la probabilité d'observer l'alignement $\mathbf{X}$, un modèle probabiliste de génération des alignements est de rigueur. Ce modèle doit être suffisamment simple pour que les calculs de vraisemblance soient possible et suffisamment flexible pour capturer les caractéristiques essentielles de l'évolution. Tous les modèles actuellement utilisés sont des chaînes de Markov à temps continu et espace d'états discrets (Swofford et al., 1996).

Modèle markovien d'évolution Deux simplifications essentielles dans la construction. On suppose tout d'abord que les sites sont indépendants. Cette hypothèse n'est pas très réaliste mais constitue une hypothèse de travail essentielle. L'indépendance simplifie grandement les calculs et sans elle, aucun modèle d'évolution ne serait utilisable. Sous l'hypothèse d'indépendance, on peut se concentrer sur l'évolution d'un seul site à la fois.

On suppose ensuite que l'évolution est Markovienne : la probabilité d'évolution d'un site ne dépend que de l'état présent de ce site et pas de ses états passés. Cette hypothèse est plus réaliste que celle d'indépendance, en effet l'évolution actuelle du génome humain, par exemple, ne dépend pas du génome des premiers primates.

Considérons $\mathcal{E}$ l'espace d'état et notons $c = |\mathcal{E}|$. Pour des séquences d'ADN, $\mathcal{E} = \{A, C, G, T\}$ et $c = 4$. Nous numérotons les états de 1 à c ($A = 1, \ldots, T = 4$). $\{1, \ldots, c\}$. L'état Y_t d'un nucléotide au temps t suit un processus de Markov à temps continu et à espace d'états $\mathcal{E}$. Le nombre de mutations, c'est à dire de sauts de la chaîne, suit un processus de Poisson homogène de taux μ. En particulier, le nombre de mutations qui surviennent en un temps t suit une loi de Poisson de paramètre μt : la probabilité que k mutations surviennent est :

$$P(k \text{ mutations}) = \frac{(\mu t)^k e^{-\mu t}}{k!}.$$

Quand une mutation intervient, nous notons $(R)_{xy}$ la probabilité de sauter d'un état x à un état y et R la matrice $R = (R_{xy})_{x,y=1..c}$. Les mutations redondantes sont autorisées, de sorte que $R_{xx} > 0$. Au final, la probabilité $P(Y_t = y | Y_0 = x)$ qu'un site saute de l'état x à l'état y en un temps t est donné par $P(Y_t = y | Y_0 = x) = P_{xy}(t)$ où $P_{xy}(t)$ est le coefficient d'indice (x, y) de

$$P(t) = \sum_{k=0}^{\infty} (R^k) \frac{(\mu t)^k e^{-\mu t}}{k!}.$$

Cette formule somme les probabilités de tous les chemins menant de x à y en un temps t.

Notons $Q = R - I$, où I est la matrice identité de taille c. On montre avec un peu d'algèbre que :

$$P(t) = \sum_{k=0}^{\infty} \frac{(R - I)^k (\mu t)^k}{k!} = \sum_{k=0}^{\infty} \frac{(Q \mu t)^k}{k!} = e^{Q \mu t}$$

La matrice Q est le générateur de la chaîne, appelé matrice de taux instantanées dans le contexte de la phylogénie. Pour $x \neq y$, $Q_{xy} dt$ est la probabilité que le site saute de x à y dans le temps infinitésimal dt.

La matrice Q joue un rôle crucial puisque tous les modèles d'évolution utilisés dans ce manuscrit sont définis en termes de contraintes sur les coefficients de Q. Le modèle d'évolution le plus simple, JC69 (Jukes and Cantor, 1969) impose par exemple l'égalité de tous les coefficients extra diagonaux. La matrice Q correspondante est :

$$Q = \begin{pmatrix} -3/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & -3/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & -3/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & -3/4 \end{pmatrix}$$

Pour lequel la matrice $P(t)$ s'écrit simplement :

$$P_{xy}(t) = \begin{cases} 1/4(1 - e^{-\mu t}), & \text{if } x \neq y \\ 1/4 + 3/4e^{-\mu t}, & \text{if } x = y \end{cases} \tag{1.1}$$

Distribution stationnaire Toujours dans l'optique de simplifier le modèle d'évolution, on impose souvent au processus Markovien d'être ergodique : quand t tend vers l'infini, la probabilité qu'un site se trouve dans l'état x converge vers une valeur non-nulle indépendante de l'état initial de ce site. Il existe des réels $\pi_A, \ldots, \pi_T$ tels que $\sum_x \pi_x = 1$ et pour tout x, y on ait $\pi_x > 0$ et

$$\lim_{t\to\infty} P_{xy}(t) = \pi_y.$$

$\pi = (\pi_A, \ldots, \pi_T)$ est la distribution stationnaire d'un site. Elle est très utile puisque pour tout $t \geq 0$

$$\pi_y = \sum_x \pi_x P_{xy}(t). \tag{1.2}$$

Autrement dit, si la distribution initiale du processus est la distribution stationnaire, la distribution de Y_t est encore la distribution stationnaire. Une fois l'ergodicité supposée, on impose donc au processus d'avoir atteint sa distribution stationnaire. Une conséquence de l'équation (1.2) est

$$0 = \sum_x \pi_x Q_{xy}$$

La distribution stationnaire peut être reconstruire directement à partir de Q. On note Π la matrice diagonale de taille c dont la diagonale vaut π. Dans notre exemple, la distribution stationnaire de JC69 est $\pi = (1/4, 1/4, 1/4, 1/4)$ est la matrice Π correspondante est :

$$\Pi = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix}$$

Chaîne réversible Toujours dans l'optique d'avoir un modèle le plus simple possible, on impose généralement à la chaîne d'être réversible : la probabilité sous la distribution stationnaire d'observer x et de sauter de x à y est égale à la probabilité d'observer y et de sauter de y à x. Autrement dit, pour $x, y \in \mathcal{E}$ et $t \geq 0$, on a :

$$\pi_x P_{xy}(t) = \pi_y P_{yx}(t)$$

qui peut se réécrire

$$\pi_x Q_{xy} = \pi_y Q_{yx}$$

ou encore ΠQ est symétrique. Le modèle JC69 est évidemment réversible. La réversibilité est intéressante dans la mesure où elle facilite le calcul de la vraisemblance ; il est en effet plus facile de calculer les valeurs propres et de diagonaliser une matrice symétrique qu'une matrice non symétrique. ΠQ étant symétrique, on montre facilement que $\Pi^{1/2} Q \Pi^{-1/2}$ l'est aussi. On peut en déduire simplement une diagonalisation de Q et le calcul de l'exponentiel de matrice $e^{Q\mu t}$ en est grandement facilité.

Un effet secondaire de la réversibilité est de supprimer la flèche du temps : la vraisemblance d'une trajectoire est indépendant du sens d'écoulement du temps. En particulier, la vraisemblance d'un arbre est indépendant de la position de sa racine (Felsenstein, 1981).

Taux de mutation En phylogénie moléculaire, les longueurs de branches ne se mesurent pas en années mais en *nombre moyen de substitutions par site*. En effet, la quantité qui intervient dans le calcul des probabilités de saut est μt, le produit du taux de mutation μ et d'une durée t. Sans information extérieure, le couple (μ, t) n'est pas identifiable, seul μt l'est. Pour pallier à cette difficulté et en l'absence d'informations extérieures, on peut faire l'hypothèse que le temps de mutation est constant le long des branches et au cours du temps pour pouvoir comparer les valeurs de μt pour différentes branches. Cette hypothèse, connue sous le nom d'horloge moléculaire, est très controversée et de plus en plus contestée (Holland et al., 2003) : les taux de mutations peuvent varier sensiblement d'une espèce à une autre.

Rappelons nous que le modèle décrit jusqu'à maintenant autorise les mutations redondantes, d'un état à lui même. Ces mutations sont une convenance mathématiques et ne devraient pas être comptées dans le taux de mutation. Sous la distribution stationnaire, la probabilité qu'une mutation soit redondante est

$$\sum_x \pi_x R_{xx} = \mathrm{Tr}(\Pi R).$$

La probabilité qu'une mutation ne soit pas redondante est donc $1 - \mathrm{Tr}(\Pi R) = -\mathrm{Tr}(\Pi Q)$ et le nombre moyen de mutations non redondantes par unité de temps est $\mu = -\mathrm{Tr}(\Pi Q)$. À cause de la non identifiabilité de μ et t et pour faciliter la comparaison des longueurs de branches entre différents modèles, on remplace Q par $\frac{1}{\mu}Q$ de sorte que le taux moyen de mutations soit 1. De cette façon, la longueur d'une branche correspond exactement au nombre moyen de mutations par site sur cette branche, et ce quel que soit le modèle.

Dans le modèle JC69 de notre exemple, le calcul donne le taux de mutation $\mu = -\mathrm{Tr}(\Pi Q) = 3/4$. Remplacer Q par version normalisée $\frac{4}{3}Q$ donne

$$P_{xy}(t) = \begin{cases} 1/4(1 - e^{-\frac{4}{3}t}), & \text{if } x \neq y \\ 1/4 + 3/4e^{-\frac{4}{3}t}, & \text{if } x = y \end{cases}$$

à comparer avec l'équation (1.1).

Lien entre modèle et contraintes sur Q Jusqu'à présent, nous n'avons introduit que le modèle JC69 qui correspond à une distribution stationnaire π uniforme et à la matrice Q très simple :

$$Q = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix}$$

normalisé de sorte que $\mu = -\mathrm{Tr}(\Pi Q) = 1$. Ce modèle particulier correspond à des contraintes très fortes sur π et Q, à savoir égalité de toutes les fréquences stationnaires et égalité de tous les coefficients extra diagonaux. Il existe cependant d'autres

paramétrisations de π et Q qui correspondent à d'autres modèles d'évolution. JC69 est le plus simple et moins réaliste de tous. De nombreuses améliorations existent, chacune correspondant à une réalité biologique non prise en compte par JC69.

Tout d'abord, les 4 nucléotides (A, C, G, T) se divisent en deux types : les purines (A, G) et les pyrimidines (C, T). Les mutations d'un nucléotide à un autre se divisent elles aussi en deux groupes : les *transitions* qui préservent le type purine/pyrimidine ($A \leftrightarrow G$ et $C \leftrightarrow T$) et les "transversions" qui modifie ($A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$ et $G \leftrightarrow T$). Les transitions n'affectent que peu les propriétés physico-chimiques des nucléotides et sont donc plus fréquentes que les transversions. Pour en tenir compte, le modèle K2P (Kimura, 1980)utilise des taux différents pour les transitions (α) et les transversions (β). Il est plus riche que JC69 : le quotient $\kappa = \alpha/\beta$ est un paramètre supplémentaire qu'on peut estimer à partir des données (bien que α et β comptent à priori pour deux paramètres supplémentaires, ils sont contraints par la normalisation $\mu = 1$). κ varie typiquement 0 et 20.

Mais on peut aussi faire tomber les contraintes sur les fréquences stationnaires. Le modèle F81 (Felsenstein, 1981) utilise des paramètres différents pour toutes les fréquences stationnaires $(\pi_A, \ldots, \pi_T)$, par opposition à JC69 qui les suppose toutes égales à 1/4. π_A, π_C, π_G sont estimés à partir des données (et π_T s'en déduit via $\pi_A + \ldots + \pi_T = 1$).

Enfin, on peut mélanger les deux. Le modèle HKY (Hasegawa et al., 1985) utilise à la fois des fréquences stationnaires inégales et un taux de transitions différent du taux de transversions. Il a 4 paramètres libres : $(\kappa, \pi_A, \pi_C, \pi_G)$. Enfin, le modèle GTR (General Time Reversible) (Lanave et al., 1984) est le modèle réversible le plus général et ne fait aucune hypothèse plus forte que l'ergodicité et la réversibilité. Les 16 Q_{xy} et les 4 π_x sont contraints par $\sum_x \pi_x = 1$ (distribution), $\pi_x Q_{xy} = \pi_y Q_{xy}$ (réversibilité), $\sum_y Q_{xy} = 0$ (matrice de taux instantanés) and $-\sum_x \pi_x Q_{xx} = 1$ (normalisation). Le modèle GTR a donc 8 paramètres et Q peut s'écrire sous la forme :

$$Q = \frac{1}{\mu} \begin{pmatrix} - & \pi_C \alpha_{AC} & \pi_G \alpha_{AG} & \pi_T \alpha_{AT} \\ \pi_A \alpha_{AC} & - & \pi_G \alpha_{CG} & \pi_T \alpha_{CT} \\ \pi_A \alpha_{AG} & \pi_C \alpha_{CG} & - & \pi_T \alpha_{GT} \\ \pi_A \alpha_{AT} & \pi_C \alpha_{CT} & \pi_T \alpha_{GT} & - \end{pmatrix}$$

où $\mu = -\sum_x \pi_x \alpha_{xx}$ et les termes diagonaux sont tels que la somme sur chaque ligne soit nulle. GTR est le plus général possible et JC69, K2P, F81 and HKY sont des cas particuliers de GTR. Les deux seuls modèles qui ne soient pas emboîtés sont F81 et K2P.

Pour résumer, le mécanisme d'évolution des données est un processus de Markov à temps continu et à espace d'état discret composé de

1. Un vecteur de probabilité $\pi = (\pi_1, \ldots, \pi_4)$: la distribution stationnaire des nucléotides. On impose $\pi_i > 0$ et $\sum \pi_i = 1$.

2. Une matrice de taux instantanés Q : pour $x \neq y$, $Q_{xy}dt$ est la probabilité de sauter de x à y en un temps infinitésimal dt. On impose de plus $Q_{xy} > 0$ pour $x \neq y$ et la contrainte de normalisation $-\mathrm{Tr}(\Pi Q)$

1.2.2 Calcul de la vraisemblance

Le modèle d'évolution présenté dans la section 1.2.1 décrit l'évolution d'un nucléotide *unique* au cours du temps, ou de façon équivalente sur une branche. Voyons maintenant comment il s'étend à un modèle d'évolution des séquences sur un arbre phylogénétique pour générer la distribution des X_i.

Nous adoptons pour cela une hypothèse d'indépendance supplémentaire : l'évolution d'un nucléotide sur des chemins (dans l'arbre) différents devient indépendante dès que les chemins se séparent, conditionnellement à l'état du nucléotide au noeud de séparation. Cette hypothèse est la suite logique des propriétés markoviennes du modèle : si l'évolution sur une branche est markovienne, l'évolution sur deux branches adjacentes ne devrait dépendre que de l'état de leur noeud commun, à l'instar de deux chaînes de Markov indépendantes démarrées au même point.

Vraisemblance d'une histoire complète Considérons un arbre phylogénétique de topologie T, de racine v_0 et de longueurs de branches $(t_e)_{e\in E}$ et un modèle markovien d'évolution, de générateur Q et de distribution stationnaire π. Un site X_i induit une fonction, elle aussi notée X_i de l'ensemble des feuilles de T dans $\mathcal{E}$. Une histoire complète $\hat{X}_i$ de X_i est une fonction de l'ensemble des noeuds dans $\mathcal{E}$ qui coïncide avec X_i sur les feuilles. Autrement dit, $\hat{X}_i$ étend X_i en attribuant un état aux noeuds internes de T.

La probabilité d'une histoire complète $\hat{X}_i$ est la probabilité du nucléotide à la racine multipliée par les probabilités de changements le long de chaque branche de l'arbre. Formellement, si on note (u, v) les deux extrémités de la branche e :

$$P(\hat{X}_i|T,(t_e),Q,\pi) = \pi_{\hat{X}_i(v_0)} \sum_{e\in E} P_{\hat{X}_i(u)\hat{X}_i(v)}(t_e) \tag{1.3}$$

Explicitons le calcul pour l'exemple simple illustré dans la Figure 1.1

$$\begin{aligned} P(A,C,C,C,G,x,y,z,w|T,(t_1,\dots,t_7),Q) = \\ P(x) \times P(y|x,t_6) \times P(A|y,t_1) \times P(C|y,t_2) \\ \times P(z|x,t_8) \times P(C|z,t_3) \\ \times P(w|z,t_7) \times P(C|w,t_4) \times P(G|w,t_5). \end{aligned}$$

Vraisemblance d'une observation L'état des noeuds internes de l'arbre est généralement inconnu la vraisemblance d'un site X_i est plus intéressante que celle d'une histoire complète $\hat{X}_i$: on ne veut pas calculer la vraisemblance conditionnellement à une certaine histoire. Pour ce faire, on somme l'équation (1.3) sur toutes les histoires complètes :

$$P(X_i|T,(t_e),Q,\pi) = \sum_{\substack{\text{Toutes les histoires} \\ \text{complètes } \hat{X}_i}} P(\hat{X}_i|T,(t_e),Q,\pi).$$

Revenons à notre exemple, le calcul de $P(X_i|T,(t_e),Q,\pi)$ nécessite une somme sur

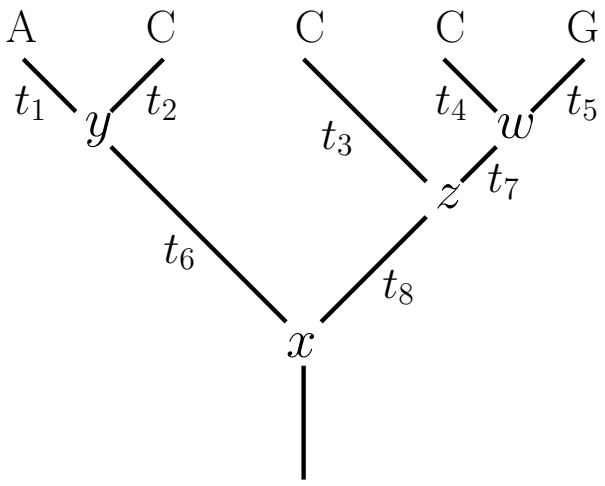

FIGURE 1.1 – Un exemple d'arbre phylogénétique. A,C,C,C et G sont les états des feuilles tandis que w, x, y, z sont les états des noeuds internes.

toutes les configurations possibles de états des noeuds internes :

$$P(X_i|T,(t_e),Q,\pi) = \sum_{x\in\mathcal{E}}\sum_{y\in\mathcal{E}}\sum_{z\in\mathcal{E}}\sum_{w\in\mathcal{E}} P(A,C,C,C,G,x,y,z,w|T,(t_e),Q,\pi)$$

Pour s espèces, il y a $s-1$ noeuds internes et le nombre de configurations croit comme c^{s-1}, de sorte que la somme n'est à priori calculable que pour de très faible valeurs de s (inférieures à 15).

Cet obstacle n'est qu'apparent, la somme peut en effet être organisée de façon de reproduire la topologie de l'arbre. Dans notre exemple,

$$\begin{aligned} P(X_i|T,(t_e),Q,\pi) = \quad & \sum_x P(x)\Big(\sum_y P(y|x,t_6)P(A|y,t_1)P(C|y,t_2)\Big) \\ & \times\Big(\sum_z P(z|x,t_8)P(C|z,t_3) \\ & \quad (\sum_w P(w|z,t_7)P(C|w,t_4)P(G|w,t_5))\Big) \end{aligned}$$

Observons que le schéma de parenthèsage des états des feuilles est $(A,C)(C,(C,G))$, qui est aussi la structure de l'arbre. Cette remarque est utilisée dans le "pruning algorithm" (Felsenstein, 1981) pour calculer la vraisemblance en $O(sc)$ opérations, nonobstant les c^{s-1} configurations.

1.2.3 Maximisation de la vraisemblance

À topologie fixée, la vraisemblance dépend des longueurs de branches et des paramètres du modèle. Pour un arbre de 30 espèces, il y a 65 tels paramètres (57 longueurs de branches et 8 paramètres pour le modèle). La maximisation de la vraisemblance est donc un problème d'optimisation non linéaire. Et il y a peu raisons de penser que la vraisemblance est convexe.

A l'exception de modèles d'évolutions très simples et d'arbres à peu de feuilles, il n'existe pas d'expression analytique des estimateurs de maximum de vraisemblance des longueurs de branches (t_e) et des paramètres (Q_{xy}) et (π_i) du modèle. Faute d'expression analytique, ces paramètres sont estimés par des méthodes d'optimisation numériques, par exemple de type Newton-Raphson.

1.3 Recherche de l'arbre

Le calcul de vraisemblance présenté dans la section 1.2.1 constitue un critère d'optimalité pour comparer différents arbres phylogénétiques. Nous présentons différentes stratégies de parcours de l'espace des arbres pour trouver le meilleur arbre.

Recherche exhaustive La première stratégie de recherche consiste à parcourir tous les arbres. C'est la plus plaisante mais la moins raisonnable. Le nombre N_s de topologies à s feuilles se calcule facilement par récurrence et vaut

$$N_s = (2s-3)!! = 1 \times 3 \times \ldots \times 2s-5 \propto s^s.$$

Pour $s = 32$ espèces, N_s vaut 8.68×10^{36}. La croissance hyper exponentielle de N_s rend impossible le parcours exhaustif des arbres, sauf pour de très petites valeurs de s. Et contrairement au calcul de la vraisemblance, il n'existe pas d'astuce qui permet de réduire le nombre d'arbres à comparer à un niveau raisonnable. La recherche de l'arbre du maximum de vraisemblance est en fait un problème *NP*-dur.

Il faut donc adopter une autre stratégie de recherche, généralement une exploration locale de l'espace des arbres. Toutes les heuristiques présentées dans la suite commencent par évaluer un arbre de départ puis se déplacent localement dans l'espace dans des arbres, en ne passant que par des arbres qui améliorent la vraisemblance avant d'atteindre un maximum local.

Heuristiques de recherche : arbre initial Il existe plusieurs façons de choisir l'arbre de départ. Nous présentons les 2 plus populaires :

1. Au hasard : l'arbre de départ est choisi au hasard parmi tous les arbres, c'est une bonne façon d'explorer l'espace des arbres mais peu d'arbres de départ sont proches sont proches du meilleur arbre.
2. Avec une autre méthode de reconstruction : l'arbre de départ est celui trouvé par une autre méthode de reconstruction, en général les méthodes de distance. Cet arbre de départ utilise mieux l'information présente dans l'alignement et est censé être plus proche du meilleur arbre qu'un arbre aléatoire.

Heuristiques de recherche : déplacements locaux Les heuristiques de recherche locale se déplacent uniquement d'un arbre de départ à un arbre d'arrivée par une transformation élémentaire de l'arbre de départ. On impose de plus que chaque soit accessible à partir d'un autre en un nombre fini de transformation élémentaire, sous peine d'être limité à un sous-ensemble strict de l'espace des arbres. Enfin, la

transformation est acceptée uniquement si elle améliore la vraisemblance de l'arbre. La recherche s'arrête quand plus aucune transformation élémentaire n'améliore la vraisemblance.

Nous présentons la transformation élémentaire la plus simple : le NNI ("nearest-neighbor interchange", échange de plus proches voisins) mais d'autres existent (Felsenstein, 2004). Le NNI agit sur un arbre en permutant deux branches adjacentes. De façon équivalente, le NNI commence par effacer une branche interne de l'arbre et toutes les branches qui lui sont adjacentes. Il en résulte 4 sous-arbres qui peuvent être réarrangés en 3 arbres différents : l'arbre de départ et deux arbres alternatifs, comme illustré en Figure 1.2. Un arbre non raciné à s feuilles comporte $s-3$ branches internes et offre donc $2(n-3)$ arbres alternatifs à l'arbre de départ. Le NNI choisit parmi ces $2(n-3)+1$ arbres celui de plus grande vraisemblance.

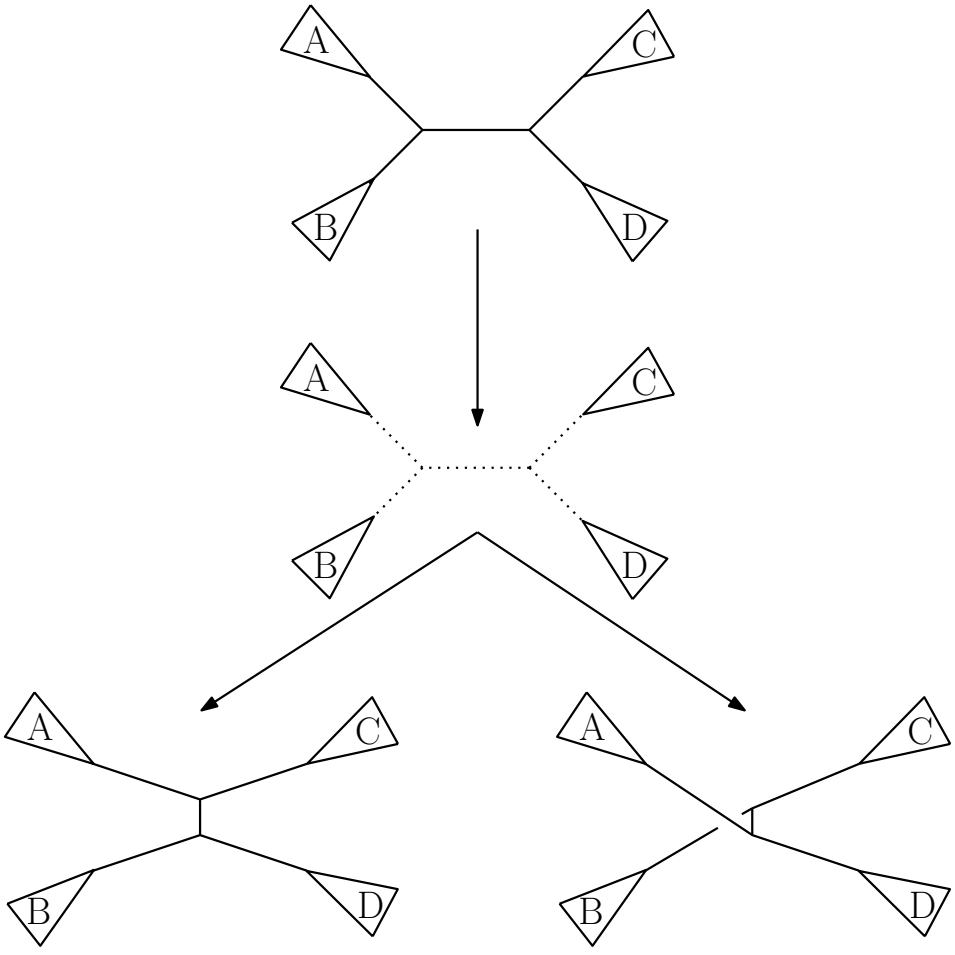

FIGURE 1.2 – Un exemple de NNI. Une branche intérieure est effacée, ainsi que les branches adjacentes et les 4 sous-arbres résultants sont réarrangés en deux arbres alternatifs.

1.4 Validation de l'arbre phylogénétique

Nous avons vu dans l'introduction qu'il est nécessaire de valider l'arbre estimé. On peut tester la topologie de l'arbre ou la valeur d'une longueur de branche. Mais on peut tester aussi considérer des tests d'hypothèses sur d'autres paramètres, comme le quotient κ du taux de transition et du taux de transversions. Certains paramètres se prêtent mieux au test que d'autres.

1.4.1 Test d'hypothèse de paramètre continu

Considérons la topologie T fixée, et un modèle d'évolution de type K2P pour lequel nous voulons tester $H_0 : \kappa = \kappa_0 = 1$ contre $H_1 : \kappa \neq 1$. Si le modèle est bien spécifié et que T est la bonne topologie, la théorie des tests (Kendall and Stuart, 1973) suggère un test de rapport de vraisemblance :

$$LR = 2 \log \frac{P(\mathbf{X}|\hat{\kappa})}{P(\mathbf{X}|\kappa = \kappa_0)}$$

Elle garantit aussi que la statistique de test LR suit asymptotiquement une loi du χ^2 à 1 degré de liberté.

$$LR \sim \chi^2_1$$

On peut utiliser cette loi asymptotique pour calibrer le test de H_0 contre H_1 au niveau voulu et construire un intervalle de confiance asymptotique.

On peut aussi tester la nullité d'une longueur de branche t_e via un test de $H_0 : t_e = t_0 = 0$ contre $H_1 : t_e > 0$. La statistique du rapport de vraisemblance est encore

$$LR = 2 \log \frac{P(\mathbf{X}|\hat{t_e})}{P(\mathbf{X}|t_0)}$$

Cependant les longueurs de branches contraintes à être positives ou nulles et t_0 est donc sur la frontière de l'espace des paramètres. Dans ce cas, LR ne suit pas asymptotiquement une loi du χ^2. Self and Liang (1987) ont prouvé que dans ce cas LR suit asymptotiquement une loi :

$$LR \sim \frac{1}{2}\delta_0 + \frac{1}{2}\chi^2_1$$

où δ_0 est la masse de Dirac en 0.

Tous ces résultats asymptotiques supposent que le modèle est correct, c'est à dire que l'évolution se comporte réellement comme une processus markovien sur un arbre et surtout que la topologie est la bonne. Mais estimer la bonne est un des buts de la phylogénie. En l'absence de certitudes sur la topologie, les tests peuvent être mal calibrés (Buckley, 2002). Enfin, les tests de rapports de vraisemblance s'appliquent bien à des paramètres continus, comme les longueurs de branches ou les coefficients de Q mais ne sont pas conçus pour des paramètres discret comme la topologie.

1.4.2 Test de topologies

Au lieu d'utiliser des tests de rapport de vraisemblance, les tests de topologies réduisent les arbres phylogénétiques à leur score de vraisemblance et compare les log-vraisemblances de topologies différentes. La vraisemblance ℓ_n^T d'un arbre phylogénétique T tend vers une valeur ℓ^T quand n tend vers l'infini. Le meilleur de deux arbres T et T' est celui de plus forte vraisemblance. Si deux arbres ont la même vraisemblance ($\ell^T = \ell^{T'}$), les différences $Z_i = \log P(X_i|T) - \log P(X_i|T')$ de log-vraisemblance en un site $Z_i = \log P(X_i|T) - \log P(X_i|T')$ sont des variables aléatoires centrées.

La majorité des tests de topologies sont construits sur cette remarque simple (Goldman et al., 2000) et testent $H_0 : E[Z] = 0$ contre $H_1 : E[Z] \neq 0$. Historiquement, le premier test est le test KH de Kishino et Hasegawa (Hasegawa and Kishino, 1989; Kishino and Hasegawa, 1989). La statistique de test est $\sum Z_i$. La distribution de cette statistique est estimée par bootstrap et permet de déterminer si $E[Z]$ est significativement différent de 0. Dans le test KH, le paramètre d'intérêt est la topologie mais le calcul des Z_i fait intervenir de nombreux paramètres de nuisance (longueurs de branches, coefficients de Q, probabilités stationnaires). Les différentes variantes du test KH se différencient par la gestion des paramètres de nuisance : estimés sur chaque nouvel alignement bootstrap ou déterminés uniquement sur l'alignement original et utilisés tels quels dans tous les alignements bootstrap. Cette dernière variante est une approximation mais permet d'accélérer la procédure de test en réduisant grandement le temps de calcul tout en donnant les mêmes résultats que la variante rigoureuse (Hasegawa and Kishino, 1994).

Cette procédure souffre cependant de deux limitations. La première est que les arbres T et T' à comparer doivent être choisis indépendemment de l'alignement. En pratique, le test KH est souvent utilisé pour comparer l'arbre du maximum de vraisemblance au deuxième meilleur arbre. Dans ce cas, $E[Z] > 0$ et le test calibré sous $E[Z] = 0$ n'est pas valide. Ensuite, le test KH est souvent utilisé pour construire un ensemble de confiance sur les arbres : plusieurs arbres sont comparés à l'arbre du maximum de vraisemblance et seuls les arbres pour lesquels le test KH ne rejette pas l'hypothèse H_0 sont inclus dans l'ensemble. Cette procédure souffre du problème de tests multiples et inclut trop peu d'arbres dans l'ensemble de confiance. Le test SH de Shimodaira and Hasegawa (1999) améliore le test KH en supprimant ces deux limitations. Sa puissance est néanmoins sévèrement limité par le nombre d'arbres à comparer ; dès que ce nombre est trop élevé, le test SH devient incapable de rejeter un arbre.

1.5 Plan de thèse

Dans la section 1.4, nous n'avons qu'abordé la question de la validation de l'arbre. Les seules méthodes présentées sont les tests de rapport de vraisemblance, pour les paramètres continus, ou assimilés, pour la topologie.

Il existe de nombreuses autres procédures de validation, en particulier des procédures non paramétrique. Certaines, comme le bootstrap et le jackknife, sont désormais classiques et présentées en détails dans le chapitre 5. Les autres constituent le coeur

de cette thèse.

Cette est organisée en deux grandes parties. Chaque partie commence par un chapitre introductif, qui introduit le sujet de recherche et en présente les enjeux. Chaque partie se poursuit avec les résultats principaux, présentés sous la forme d'articles de recherche, acceptés ou soumis à des revues à comité de lecture et s'achève avec une discussion des résultats et de nouvelles perspectives de recherche.

1.5.1 Variabilité normale du modèle

Variabilité d'échantillonnage Nous considérons dans la partie I la variabilité normale de l'estimateur, inhérente à la nature probabiliste du modèle. Le premier type de variabilité est la variabilité d'échantillonnage. Pour un modèle markovien $M = (Q, \pi)$ donné d'évolution des nucléotides, on peut associer à chaque arbre phylogénétique $T = (T, (t_e))$ sa log-vraisemblance asymptotique ℓ^T. Les arbres sont classés du pire au meilleur par valeurs croissantes de ℓ^T et l'arbre du maximum de vraisemblance est exactement celui qui maximise ℓ^T. La valeur ℓ^T est cependant inaccessible, à moins de disposer d'alignements de longueur infinie ou de connaître parfaitement la distribution Q des sites. Faute de mieux, ℓ^T est estimé par ℓ_n^T, la vraisemblance de T calculé sur un n-échantillon de Q. ℓ_n^T est une variable aléatoire centrée sur ℓ^T mais de variance non-nulle ; les classements d'arbres induits par ℓ_n^T et ℓ^T ne sont pas forcément identiques. En particulier, deux arbres T et T' peuvent vérifier $\ell_n^T < \ell_n^{T'}$ bien que $\ell^T > \ell^{T'}$. Autrement dit, l'arbre sélectionné sur seulement n observations n'est pas forcément le meilleur des deux. Dans le chapitre A, nous utilisons des inégalités de concentrations pour contrôler les fluctuations de ℓ_n^T autour de ℓ^T. Nous bornons aussi avec les mêmes techniques la probabilité de choisir entre deux arbres celui de plus basse vraisemblance à cause de la variabilité d'échantillonnage.

Détection de point de rupture Un second type survient quand la loi des sites change le long de la séquence. Sous l'hypothèse que les sites sont indépendants et de même loi Q, le score de vraisemblance ℓ_n^T converge vers sa valeur asymptotique ℓ^T quand n tend vers l'infini. Il est donc raisonnable d'utiliser des alignements aussi long que possible pour estimer ℓ^T. Cependant, si la loi des sites saute de Q à Q' au delà d'un certain nombre d'observations n_0, la valeur limite de ℓ_n^T change aussi. Dans ce cas là, utiliser les observations d'indice supérieur à n_0 perturbe l'estimation de ℓ^T et au final de T. Avant d'utiliser un nouveau paquet d'observations, par exemple un gène nouvellement séquencé, pour améliorer l'estimation de T, il est nécessaire de s'assurer qu'elles ont la même loi que les observations précédentes. Estimer ℓ^T revient à estimer la moyenne d'une distribution donnée. Nous développons dans le chapitre B un test non paramétrique pour détecter des sauts dans la moyenne de données séquentielles. Notre test est construit sur l'intuition que les variations typiques de la moyenne résultant de l'ajout de nouvelles observations au jeu de données sont équivalentes aux variations typiques résultants du retrait du même nombre d'observations. Cette intuition est formalisée par des développements de Edgeworth, qui fournissent de plus les termes correctifs au premier ordre. Ces termes sont valables pour des observations continues aussi bien que discrètes.

1.5.2 Détection des données aberrantes

Sites influents Nous nous intéressons dans la partie II à la variabilité induite par des données aberrantes. La méthode d'estimation doit non seulement converger vers le bon arbre avec la variabilité d'estimation la plus faible possible mais elle doit aussi être robuste. En effet, si de petites modifications de l'alignement (par exemple des erreurs d'alignement ou de séquençage) ont une influence drastique sur l'arbre estimé, la fiabilité de cet arbre est sujette à caution. La méthode la plus simple de construire un arbre robuste est sans doute de se débarrasser des données aberrantes, ou au moins de leur attribuer une importance réduite dans la procédure d'estimation. Mais pour ce faire, il faut connaître les données aberrantes. Dans le chapitre C, nous adaptons les outils traditionnels de l'analyse de robustesse (fonction d'influence et courbe de sensibilité) à la phylogénie pour détecter les sites influents. Nous les appliquons ensuite à un jeu de donnée de mycètes dans lequel nous identifions avec succès des sites atypiques, qui perturbent fortement la reconstruction de l'arbre.

Espèces influentes Les sites de l'alignement jouent un rôle crucial puisqu'ils fournissent l'information nécessaire à la reconstruction de l'arbre phylogénétique des espèces inclues dans l'alignement. Mais les espèces ne sont pas en reste : suivant les espèces inclues dans l'alignement, l'arbre sera plus ou moins facile à reconstruire. Plus surprenant, l'échantillonnage d'espèces a lui aussi un impact sur la topologie reconstruite. Là encore, si un arbre est grandement modifié par l'ajout ou le retrait de quelques espèces à l'alignement, sa fiabilité doit être mise en question. La construction d'arbres robustes à l'échantillonnage d'espèces nécessite l'identification préliminaire des espèces influentes. Une façon naturelle de mesurer l'influence d'une espèce sur un arbre à l'aide du jackknife. Le jackknife d'espèces est un problème statistique inhabituel puisque les espèces, contrairement aux sites, ne sont pas indépendantes. Nous proposons néanmoins dans le chapitre D une adaptation des fonctions d'influence aux espèces. Nous l'appliquons ensuite à un jeu de données de mammifères à placenta dans lequel nous retrouvons des espèces bien identifiées dans la littérature comme des espèces influentes.

Chapter 2

Introduction to Phylogenetics

In this part, we give an overview of molecular phylogenetics and introduce the main topic of this thesis which the issue of variability and robustness in phylogenetics. We start with a historical presentation and motivation of molecular phylogenetics (sec. 2.1). We focus on the maximum likelihood method, which requires us to design a probabilistic model for evolution (sec. 2.2) and to search for the estimate of a tree (sec. 2.3). In return, it provides a relevant framework to study the variability of the phylogenetic estimates. We then present the peculiar characteristics of the variability issue in phylogenetics and review likelihood-based tests to assess the validity of a tree (sec. 2.4). We conclude this chapter by giving the outline of this thesis in which we propose original methods to quantify the variability of a tree, detect outliers and propose robust estimates of the tree.

2.1 A bit of Context

2.1.1 A Brief History of Molecular Phylogenetics

Charles Darwin's *On the Origin of Species* (Darwin, 1859) is the seminal work considered to be the foundation of modern evolutionary biology. Darwin's work introduced the theory of evolution, that populations evolve over the course of generations through the process of natural selection, and the concept that the diversity of life arose through a branching pattern of evolution and common descent.

Darwin's natural selection can be simply stated: "Heritable traits that increase reproductive success will become more common in a population". Thus, in order for selection to act, there must be variation within a population and offspring must be similar to their parents. One difficulty at the time was that Darwin's book gave no credible explanation about what served as a support for the traits. The other is that variation within a population should disappear under the influence of selection and Darwin gave no credible source of variation within a population. Both are provided by Mendelian genetics (Mendel, 1866). The idea can be simply stated: "traits are determined by genes". Each gene occurs in different types called alleles and different alleles may produce different traits. "Mutations" can cause the apparition of new

alleles and maintain variation within a population. Darwin argued that evolution of complex, well-adapted organisms depends on selection acting on a large number of slight variants of a trait whereas much of Mendel's work deliberately focused on discontinuous changes in traits determined by a single gene. The two views look incompatible at first.

The resolution of this incompatibility lays in the pioneering work of Fisher, Haldane and Wright. In 1918, Fisher showed that Mendelian genetics are consistent with natural selection (Fisher, 1918). If traits depend on multiple genes, each making a small contribution, the discontinuous nature of Mendelian alleles is reconciled with continuous variation and gradual evolution. Starting in 1924, Haldane studied how differences in survival or reproduction of one or two Mendelian genes would affect the populations, showing that evolution can act extremely fast in real world examples, such as peppered moths (see Larson (2006) for a complete account). Starting in 1921, Wright quantified the way the random process of reproduction in a finite population would lead to changes in allele frequency and examines how it interacts with selection and mutation. The work of Fisher, Haldane and Wright founded population genetics. It was the stimulus for the modern evolutionary synthesis which gives a logical account of evolution, by drawing ideas from several branches of biology.

At the time of the evolutionary synthesis, genetic variability could not be observed. Early work was restricted to genes that happened to be detectable in some ways, but things changed dramatically over the subsequent fifty years. In 1953, Watson and Crick (Watson and Crick, 1953) published and described the discovery of the double helix structure DNA. Doing so, they solved a fundamental mystery about living organisms and revealed how genetic instructions are stored inside organisms and passed from generation to generation. In response, researchers from molecular biology, evolutionary biology and population biology sought to understand recent discoveries on the structure and the function of DNA and protein. They also turned to study evolution at the scale of DNA, RNA and proteins. This consecrated molecular evolution as a scientific field in the 1960s.

The rise of molecular evolution considerably improved our understanding of evolution. Kimura introduced in 1968 the neutral theory of evolution (Kimura, 1968), which states that the vast majority of evolutionary changes at the molecular level are caused by random drift and neutral mutations. This challenged the up to then prevailing view that selection and adaptive process were the dominant forces of evolution.

The ability to sequence and analyze biological macromolecules (RNA, DNA, proteins) drastically changed not only the prevailing view of evolution but also the traditional field of systematics and gave birth to phylogenetics. Macromolecules provide evidence of descent at a finer and more reliable level than phenotypic traits, and permits us to work out the evolutionary relationships among various species believed to have a common ancestor.

Evolution is thought of as a branching process whereby populations are altered over time, may speciate into separate branches or terminate, hence the visualization by an evolutionary "tree". Darwin first illustrated and popularized this notion in his book. Notably, the tree diagram used to show the divergence of species is the only illustration in *On the Origins of Species*. Trees diagrams (phylogenetic trees) are still

in use to depict evolution. The tree visualization conveys the concept that speciation occurs through the adaptive and random splitting of lineages.

2.1.2 Molecular Phylogenetics in Biology

> *Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species, and to analyze those differences statistically.* (Felsenstein, 2004)

Molecular phylogenetics is a lively field of research with a number of practical applications. As the product of systematics and molecular genetic, its most emblematic role is certainly on reconstructing the Tree of Life, which describes the relationships of all life on Earth from an evolutionary point of view (Tree of Life Web Project at `www.tolweb.org` and Dunn et al. (2008) for the animal tree of life). The prospect of reconstructing the Tree of Life is intrinsically appealing to evolutionary biologists but phylogenetics also has a wide range of applications throughout biology.

Phylogenies are priceless in comparative genomics. They allow us to account for the non-independence of organisms when analyzing genomes across several biological levels (for example, genes, genomes, individuals, species, etc). They allow biologists to predict the function of a unknown gene from its function in closely related species (Eisen, 1998; Eisen and Wu, 2002; Gu et al., 2005; Yu et al., 2004). They can be used in protein-protein interaction networks to infer edges, that is interaction between proteins (Pazos and Valencia, 2001; Wuchty, 2004). Phylogenies are also useful to study gene families expression profiles while accounting for non-independence induced by the phylogeny of the genes (Gu, 2004; Guo et al., 2007). Without a phylogenetic framework, each genome would be independent from every other, comparative analysis would be meaningless and detection of a gene function would be greatly hindered. The study of human genes function would require human experimentation, which is not ethical, since results on primate genes could not be transposed to or give a hint about human genes. The phylogenetic framework has therefore very practical implications about our understanding of how humans function.

But phylogenetics is by no means limited to the study of genomes. It can also be used in human health to predict the evolution of Human Influenza (Bush et al., 1999), in conservation biology to select the best way to preserve biodiversity when only a limited number of taxa can be conserved (Bordewich et al., 2008), in taxonomy to identify from a DNA "barcode" (a short genetic marker of mitochondrial DNA) the species of an organism (Hebert et al., 2003). More exotic applications include forensics where DNA profiling, also known as genetic fingerprinting, can identify individuals based on their DNA profile (Jeffreys et al., 1985).

More generally, phylogenies are instrumental in some of the hot topics of molecular evolution, such as the role of gene duplication in the emergence of novel gene function or the extent of adaptive versus neutral evolution.

2.1.3 Methods for Inferring Trees

> *I am very pleased to see that the problem offers sufficient challenges to statisticians...*
> Luca Cavalli-Sforza in Edwards (1970)

All applications of phylogenetics described in Section 2.1.2 require accurate phylogenetic estimates but inferring a phylogeny is not an easy task. First, evolution is a unique event, the process of which is impossible to observe or repeat. And genetic information is available only for extant taxa, not for ancestral ones which have gone extinct for quite some time. Phylogenetics is about inferring the path evolution took through the footprints it left on extant species genomes, no matter how faint the footprints may be or how quickly they may vanish. This is a daunting task and yet, phylogenies have been inferred by systematics since Darwin. We are primarily interested in algorithmic methods, that can be carried out by a computer.

The two contenders for the title of first paper on numerical inference of phylogenies are Michener and Sokal (1957) and Edwards and Cavalli-Sforza (1963). In Michener and Sokal (1957), the authors developed clustering methods for biological classifications. The purpose, stated by Michener, was not simply to classify but also to infer the phylogeny. The creative work of Edwards and Cavalli-Sforza, both students of Fisher, lies at the foundation of numerical work on phylogenies. In their 1963 abstract, they stated the parsimony method for the first time. In Edwards and Cavalli-Sforza (1964), they presented the parsimony method, the maximum likelihood method and the statistical approach to inferring phylogenies. In Cavalli-Sforza and Edwards (1967), they introduced least-square methods to the field. Their most important contribution is perhaps the early realization that phylogeny reconstruction is a statistical inference problem although Farris (1983) questioned it. Later on, Fitch and Margoliash (1967) introduced and popularized distance matrix methods. Felsenstein proposed an algorithm for the application of maximum likelihood to discrete characters (Felsenstein, 1973) and the 'pruning algorithm' to relax some of maximum-likelihood computational limitations (Felsenstein, 1981). Mau (1996) and Li (1996) in their Ph.D. thesis and Rannala and Yang (1996) in a article described bayesian phylogenetic inference, although it was already mentioned in Felsenstein's Ph.D. thesis (Felsenstein, 1968).

All previously mentioned methods for inferring phylogenies are made up of three parts:

1. an *optimality criterion* to measure how the data fit a particular tree (e.g. parsimony, likelihood, sum of square, etc);
2. a *search strategy* for finding the optimal tree(s);
3. assumptions about the *mechanisms of evolution* for the data.

The problem of choosing a method has by no means been "solved"; debates continue to rage about the merits of competing methods. Sometimes the choice of a method does not matter; the same phylogeny is found by all methods. This is of the course the ideal situation, rarely encountered in practice. Penny et al. (1992) suggest that a method should have five desirable properties: *scalability*, *consistency*, *efficiency*, *robustness* and *falsifiability*. Scalability means the method is not overly resource-demanding and can be used to analyze medium to large data sets. Consistency means

that the method converges to the correct tree when the number of observations goes to infinity. Efficiency means that the method has the smallest possible variance around the correct true so that convergence can be achieved with relatively small number of observations. Robustness means that the method remains consistent under small deviations from the model, when the assumptions about the mechanisms of evolution are slightly off. Finally the data must be able, in principle, to falsify the model. This is perhaps the most difficult feature.

2.1.4 Validating the tree

This thesis presents some methods to quantify the variability of the inferred tree. Up to now, no method has been proved to possess the five desirable properties. Nevertheless, some proved very useful and greatly improved our understanding of evolution. In this thesis, we concentrate on phylogenetic methods based upon the likelihood function – the method of maximum likelihood and bayesian inference. We believe that likelihood-based methods have the greatest potential to possess some if not all of the properties. Maximum likelihood and bayesian inference differ in the way they envision probabilities and inference problem but make the same use of likelihood function, which is to carry the information about phylogeny contained in the data. For a comprehensive description of other methods such as parsimony and distance methods we refer to Felsenstein (2004).

All inference methods share the same concern; the inferred tree needs to be validated in some way. Consistency is of course a desirable property but provides no guarantee whatsoever that the inferred tree is the correct one. The inferred tree should therefore be evaluated to see how similar it is to the correct tree. The simplest way to do this would be to compare the inferred tree to the correct one to see how different they are, using for example tree metrics. Unfortunately, if the correct tree was known, we would not bother trying to infer it.

Likelihood based estimation of a phylogeny is a statistical inference problem; the estimate of a phylogeny is associated with some variability and the statistical framework is useful to quantify this variability. The natural move, inspired by asymptotic statitics, would be to prove a convergence theorem (usually asymptotic normality) for some normalization of the distance between the inferred and the correct tree. Unfortunately, phylogenetic inference is a peculiar statistical problem (Yang et al., 1995) as the tree is made up of continuous branch lengths but discrete topology. Furthermore the tree space is not embeddable in an euclidian space (Billera et al., 2001) and the usual techniques to prove asymptotic normality prove useless, although some attempts exists (Holmes, 2005). Quantifying the variability of the inferred tree is not easily done with limit theorems and people usually resort to resampling methods such as bootstrap (see Sec. 2.4 and Chapter 5) or jackknife (see Chapter 5 and C). We can also focus on other features of the tree, if possible embeddable in a $\mathbb{R}^d$ for some d. Finally, the variability may be induced by sources that the statistical framework does not handle well. They should be quantified nevertheless.

This thesis presents some methods to quantify this variability. More specifically, we focus on pinpointing and quantifying the variability induced by specific sources: observation sampling, inconsistency of the phylogenetic signal along the sequence,

outlier sites and influent species. The rest of Chapter 2 presents the likelihood function (Sec. 2.2), describes some tree search strategies (Sec. 2.3) and some validation procedures (Sec. 2.4). Finally, section 2.5 gives the outline of this thesis.

2.2 Likelihood function in Molecular Phylogenetics

This section borrows heavily from Huelsenbeck and Bollback (2007), Bryant et al. (2005) and Chapter 20 of Felsenstein (2004). Likelihood requires essentially data and a model of how the data arose. This model gives a probability $P(D|\theta)$ of observing the data, given value θ for the parameters of the model. In phylogenetics, θ includes a tree, branch lengths, a model of sequence evolution and many others. The key idea behind likelihood is to choose the parameters that maximize the probability of observing the data that were actually observed. To do this, we define the likelihood function $L(\theta) = P(D|\theta)$ that captures how "likely" it is to observe the data for a given value θ of the parameter. A high likelihood indicates a good fit. The *maximum likelihood estimate* (ML estimate) maximizes $L(\theta)$. In the context of phylogenetics, we search for the ML estimate of a phylogeny.

The basic intuition behind likelihood inference is surprisingly straightforward but its application to phylogenetics is quite difficult. The first problem is model design; what model should we use for evolution ? The second problem is computational; can we effectively compute the likelihood and optimize the model parameters ? The third and last problem is validation and significance; are the results significant and reliable ?

2.2.1 Three parts of an evolutionary model

Models used to describe the data can be broken down in three components:
- a topology;
- a mechanism of change (usually Markov model of sequence evolution);
- parameters needed to specify the model.

Terminology for graph and phylogenetic trees A graph $G = (V, E)$ is a set V of nodes (or species) and a set $E \subset V \times V$ of edges (or branches). If edge e is represented by (v_1, v_2), e is incident on v_1 and v_2. Two edges are adjacent if they share a common node. A path is a sequence of edges $e_1, \ldots, e_n$ where e_1 is adjacent to e_2,..., e_{n-1} is adjacent to e_n. If $e_1 = (v_0, v_1)$ and $e_n = (v_{n-1}, v_n)$, we say v_0 and v_n are the endpoints of the path. A graph is connected if for each pair of nodes, there is a path connecting them (see Fig. 2.1). A cycle is a path with two ends being the same point. The degree of a node is the number of edges incident to that node. A node of degree 1 is a leaf.

A tree is a graph which is connected and has no cycles: a "connected acyclic graph". For phylogenetic trees, a node represent a taxon by attaching a label to the node. Often the edges (branches) are associated with a real number, the weight (or length) of the edge. In phylogenetics, this can represent a number of substitutions, a time or a probability of substitution; hence weights are non-negative. We often wish

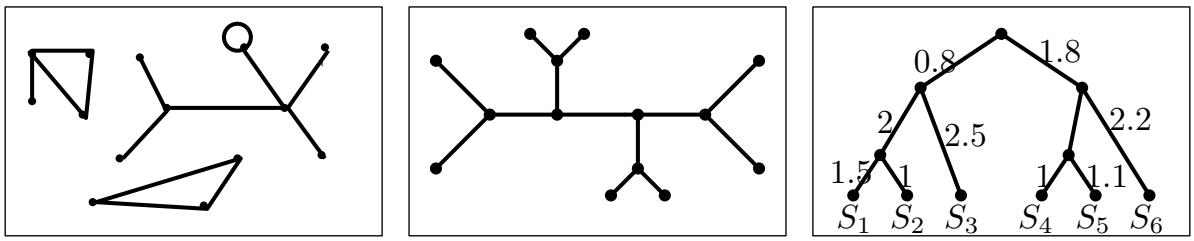

Figure 2.1: Illustrations of unconnected graph, unrooted tree and rooted weighted tree.

to identify the root of the tree (the most recent common ancestor). Such a tree is a rooted directed tree; the root is unique and all edges are directed away from the tree. A rooted tree can be derived from an unrooted tree by identifying a root, which may be a existing node or a new node (of degree 2).

A phylogenetic tree for a set of taxa S is a tree with labels on all leaves. A phylogenetic tree may be rooted or unrooted, weighted or unweighted binary or non binary. In a rooted tree, a path from an internal node to a leaf is called a lineage; it is directed away from the root. A subtree is the set of lineages from a node v with a common first edge e. In a binary rooted tree, any path from the root has two choices (a bifurcation) at every non-leaf node. A topology is an unweighted phylogenetic tree.

Character State Matrix Before turning to the mechanism or change, we give some consideration to the data. For the phylogeny problem, the observations are taken to be the aligned character matrix. For DNA or amino acids sequence data, the alignment will be denoted **X**. An example of character matrix (or data matrix, data set or simply data) is given in Table 2.1. Each row in the matrix is a sequence and represents the information about a taxon. A taxon can be an individual, a group, a population, a species or some higher taxonomic group. Each column of the matrix is a site (also called column, position, nucleotide, amino acid) or a character. The individual observations are the columns, denoted $X_1, \ldots, X_n$. In our example, $X_4 = (N, N, N, N, N, D)$.

Multiple sequence alignment (MSA) is not an easy task. When aligning two sequences, dynamic programming guarantees a mathematically optimal alignment. However, attempts at generalizing dynamic programming to multiple alignments are limited to small numbers of short sequences. Any alignment for more than eight genes or proteins of moderate length is untractable and practical alignments methods, such as the popular ClustalW (Thompson et al., 1994), make use of heuristics. The heuristics are usually based on homologous sequences being evolutionary related. The alignment is progressively made by a series of pairwise alignment following the branching order of a phylogenetic tree; the most closely related sequences are aligned first, the more distant later. Alignment and phylogenetic inference rely on each other and should ideally be performed together (Liu et al., 2009). We do not address this issue here and consider only aligned data.

Fin Whale	M	N	E	**N**	L	F	A	P	F
Blue Whale	M	N	E	**N**	L	F	A	P	F
Chimpanzee	M	N	E	**N**	L	F	A	S	F
Bonobo	M	N	E	**N**	L	F	A	S	F
Gorilla	M	N	E	**N**	L	F	A	S	F
Bornean Orangutan	M	N	E	**D**	L	F	T	P	F

Table 2.1: : Example of a character matrix of size $s \times n$ (6 species $\times$ 10 characters). X_4 is highlighted.

2.2.2 Markov models of sequence evolution

In order to compute the probability of observing $\mathbf{X}$, we need a probabilistic model for evolution. Evolution is of course so complex that no model can be completely accurate and we have to make simplifying assumptions. Currently, all models of evolutionary change on a phylogeny are continuous-time discrete state Markov models. A complete review and discussion of the models can be found in Swofford et al. (1996).

The first simplification is the assumption that sites evolve independently. It is a bit unrealistic but still an essential working assumption. Independence greatly simplifies computations and without it, no model would be tractable. With the assumption of independence, we can focus on the evolution of single site.

The second assumption we make is that evolution is Markovian. It requires that the probability of substitution depends only on the present nucleotide at a site and not of what the sequence was earlier in evolution. This is realistic as no informaton is conserved in, for example, the human sequence as to which nucleotide was present in early primates.

Basic Model Let $\mathcal{E}$ bet the space of states and $c = |\mathcal{E}|$. For DNA sequences $\mathcal{E} = \{A, C, T, G\}$ and $c = 4$ while for proteins, $\mathcal{E}$ is the set of amino acids and $c = 20$. We index the states in $\{1, \ldots, c\}$. The mutation events occur according to a *continuous time Markov chain* with state set $\mathcal{E}$ and rate μ. The number of mutations in time t has Poisson distribution with parameter μt; the probability of k mutation events is

$$P(k \text{ events}) = \frac{(\mu t)^k e^{-\mu t}}{k!}.$$

When a mutation event happens, we note R_{xy} the probability of changing from state x to state y and R the matrix $R = (R_{xy})_{x,y}$. Since redundant mutations are allowed, $R_{xx} > 0$. Wrapping everything together, the probability $P(Y_t = y | Y_0 = x)$ that a site turns change from state x to state y in time t is given by $P(Y_t = y | Y_0 = x) = P_{xy}(t)$ where $P_{xy}(t)$ is the xyth coordinate of

$$P(t) = \sum_{k=0}^{\infty} (R^k) \frac{(\mu t)^k e^{-\mu t}}{k!}.$$

This formula expresses the probabilities of change summed up over all possible paths from x to y in time t.

Noting $Q = R - I$, where I is the $c \times c$ identity matrix, a bit of matrix algebra gives

$$P(t) = \sum_{k=0}^{\infty} \frac{(R-I)^k(\mu t)^k}{k!} = \sum_{k=0}^{\infty} \frac{(Q\mu t)^k}{k!} = e^{Q\mu t}$$

The matrix Q is called the instantaneous rate matrix and most models of sequence evolution are defined in terms of their instantaneous rate matrix. For $x \neq y$, $Q_{xy}dt$ is the probability that a site change from x to y in a (infinitely) small time dt.

Although $e^{Q\mu t}$ looks complicated, there is a standard trick to compute it. If Q can be diagonalized as $Q = ADA^{-1}$ with D diagonal, then

$$e^{Q\mu t} = Ae^{D\mu t}A^{-1}$$

where e^D is a diagonal matrix with $(e^D)_{xx} = e^{D_{xx}}$.

Consider the JC69 (Jukes and Cantor, 1969) as an example. We assume that the states are ordered A, C, G, T. The model is defined by the (instantaneous) rate matrix:

$$Q = \begin{pmatrix} -3/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & -3/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & -3/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & -3/4 \end{pmatrix}$$

Rows in Q correspond to the initial state and columns to the final state. This model is equivalent to one with discrete generations occurring according to a Poisson process and transition probability matrix

$$R = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

Diagonalizing Q and taking the exponential to obtain the transition probability matrix $P(t)$ simply gives

$$P_{xy}(t) = \begin{cases} 1/4(1 - e^{-\mu t}), & \text{if } x \neq y \\ 1/4 + 3/4e^{-\mu t}, & \text{if } x = y \end{cases} \tag{2.1}$$

Stationary Distribution We often assume that the Markov process is ergodic; as t goes to infinity, the probability that a site is in state x converges to a value which is non-zero and independent of the starting site. There exist positive $\pi_1, \ldots, \pi_c$ summing up to 1 such that, for all x, y:

$$\lim_{t\to\infty} P_{xy}(t) = \pi_y.$$

$(\pi_1, \ldots, \pi_c)$ is the *stationary distribution* of the states. They are convenient because for all $t \geq 0$

$$\pi_y = \sum_x \pi_x P_{xy}(t). \tag{2.2}$$

If we sample from the stationary distribution and run the process for any time, the final state is also at the stationary distribution. Therefore, we usually assume that the process reached its stationary state. A consequence of equation (2.2) is

$$0 = \sum_x \pi_x Q_{xy}$$

We can recover the stationary distribution directly from Q. We use Π to note the $c \times c$ diagonal with (π_x) down the diagonal. From equation (2.1), we see that the stationary distribution of JC69 is $(1/4, 1/4, 1/4, 1/4)$. The matrix Π is

$$\Pi = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 0 & 0 \\ 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \end{pmatrix}$$

Time Reversibility The next common assumption is that of *time reversibility*. We assume that the probability of sampling x from the stationary distribution and going to y is the same as the probability of sampling y from the stationary distribution and going to x. That is, for $x, y \in \mathcal{E}$ and $t \geq 0$ we have

$$\pi_x P_{xy}(t) = \pi_y P_{yx}(t)$$

which corresponds to the condition that

$$\pi_x Q_{xy} = \pi_y Q_{yx}$$

or in other words, ΠQ is symmetric. JC69 is obviously time reversible. Time reversibility makes it easier to diagonalize Q because it is, in general, easier to find eigenvalues of a symmetric matrix than of a non symmetric matrix. Since ΠQ is symmetric, so is $\Pi^{1/2} Q \Pi^{-1/2}$. We first diagonalize $\Pi^{1/2} Q \Pi^{-1/2}$ as

$$\Pi^{1/2} Q \Pi^{-1/2} = BDB^{-1}$$

where D is diagonal and B invertible. Setting $A = \Pi^{-1/2} B$ gives $Q = ADA^{-1}$. This is the approach used by most inference software when computing the exponential of general rate matrix for which explicit formula for $e^{Q\mu t}$ are not available in general. A welcome by-product of time reversibility is to make computation of likelihood easier. Since can flow both directions, the likelihood becomes independent of the position of the root (Felsenstein, 1981).

Mutation Rate In molecular phylogenetics, time and branch lengths are measured in *expected number of mutations per site* rather than years. The reason is that the relevant measure of the quantity of evolution is μt, the product of mutation rate μ and (real) time t. Without exterior information, one must assume equal mutation rates μ in order to compare times t across branches. This is known as the molecular clock hypothesis and is a highly controversial hypothesis with vanishing support (Holland et al., 2003); mutation rates can markedly change between species and branch lengths are highly species dependent.

Recall that our model of site evolution has mutations events occuring according to a Poisson process with rate μ over a period t, with expected number of event μt. However, redundant mutations, from a site to itself, are a mathematical convenience and should not be counted in the mutation rate. Sampling from the stationary distribution, the probability that a mutation is redundant is

$$\sum_x \pi_x R_{xx} = \mathrm{Tr}(\Pi R).$$

Hence the probability that a mutation is not redundant is $1 - \mathrm{Tr}(\Pi R) = -\mathrm{Tr}(\Pi Q)$ and the expected number of (non-redundant) events in a unit time ($t = 1$) is $\mu = -\mathrm{Tr}(\Pi Q)$. To facilitate comparison between models, we replace Q by the normalized $\frac{1}{\mu}Q$ so that the overall rate of mutation is 1. In this way, the length of a branch corresponds to the expected number of mutation on that branch, irrespective the model.

For the JC69 model we obtain a rate of $\mu = -\mathrm{Tr}(\Pi Q) = 3/4$ so that we replace Q by $\frac{4}{3}Q$ to normalize the rate. This leads to

$$P_{xy}(t) = \begin{cases} 1/4(1 - e^{-\frac{4}{3}t}), & \text{if } x \neq y \\ 1/4 + 3/4e^{-\frac{4}{3}t}, & \text{if } x = y \end{cases}$$

to compare with equation (2.1).

Discussing the models and the different shapes of Q The only model introduced by now is the JC69 model which corresponds to the very simple instantaneous rate matrix

$$Q = \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix}$$

normalized so that $\mu = -\mathrm{Tr}(\Pi Q) = 1$. There are however many other models for DNA sequences, each based on a different parametrization of Q. JC69 is the easiest and less realistic one. Several refinements have been proposed, that capture different features of sequence evolution. The 4 nucleotides (A, C, G, T) belong to two categories: Purine (A, G) or Pyrimidine (C, T). Nucleotide substitutions can be divided in two types: "transitions" that conserve the purine/pyrimidine status ($A \leftrightarrow G$ and $C \leftrightarrow T$) and "transversions" that change it ($A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$ and $G \leftrightarrow T$). Transitions correspond only to slight modifications of the properties of the nucleotides and happen much more frequently that transversions. The first refinement, K2P introduced by Kimura (1980), assumes different rates for transitions (α) and transversions (β). It is more complex that JC69; the ratio $\kappa = \alpha/\beta$ is a free parameter estimated from the data (α is constrained by the normalization $\mu = 1$). κ usually ranges between 0 and 20. An independent refinement, F81, was introduced by Felsenstein (1981). It assumes free nucleotides frequencies $(\pi_A, \ldots, \pi_T)$, in opposition to JC69 which assumes $\pi_A = \ldots = \pi_T = 1/4$. π_A, π_C, π_G are free parameters estimated from the data (π_T is deduced from them by $\pi_A + \ldots + \pi_T = 1$). The HKY model (Hasegawa et al., 1985) assumes both unequal stationary nucleotide frequencies and different rates for transitions and transversions. It has 4 free parameters $(\kappa, \pi_A, \pi_C, \pi_G)$. Finally the GTR (General Time Reversible), introduced by Lanave et al. (1984), does not assume anything. It has 20 parameters, 16 Q_{xy} and 4 π_x, constrained by $\sum_x \pi_x = 1$ (frequencies), $\pi_x Q_{xy} = \pi_y Q_{xy}$ (reversibility), $\sum_y Q_{xy} = 0$ (instantaneous rate matrix) and $-\sum_x \pi_x Q_{xx} = 1$ (normalization). There are thus 8 free parameters and Q can be parametrized

$$Q = \frac{1}{\mu}\begin{pmatrix} - & \pi_C\alpha_{AC} & \pi_G\alpha_{AG} & \pi_T\alpha_{AT} \\ \pi_A\alpha_{AC} & - & \pi_G\alpha_{CG} & \pi_T\alpha_{CT} \\ \pi_A\alpha_{AG} & \pi_C\alpha_{CG} & - & \pi_T\alpha_{GT} \\ \pi_A\alpha_{AT} & \pi_C\alpha_{CT} & \pi_T\alpha_{GT} & - \end{pmatrix}$$

where $\mu = -\sum_x \pi_x \alpha_{xx}$ and the diagonal terms are such that the rows sum up to zero. This is the most general time reversible DNA model and JC69, K2P, F81 and HKY are special case of GTR. The only two models not nested one in the other are F81 and K2P.

The story is different for protein sequences. Since instantaneous rate matrices are 20×20, they are not estimated from the data while inferring the phylogeny. Q is rather an empirical rate matrix, estimated on the same or another data set, and used as such during the inference (Jones et al., 1992).

2.2.3 Computing the Likelihood

We are now completely prepared to compute the likelihood of a site on a tree. Section 2.2.2 discuss Markov model describing the evolution of a *single* nucleotide along time, or equivalently on a single branch. We now extend the model for sequence evolution to evolution on a phylogeny. We do not need to work with full sequence. Again, since sites evolve independently, studying a single site is enough. We want however to describe evolution on many branches (a tree) rather than on a single one.

To do this we need to make one more independence assumption. We assume that evolution in different lineages is independent; evolution on a branch depends only on the state at the beginning of that branch. This is reasonable enough. The Markovian properties ensure that evolution of a site on a branch depends only on the state of that site at the beginning of branch. Evolutions of the same site on adjacent branches, given the state of that site at their common node, should be independent, just like independent Markov chain started from the same point run independently.

The likelihood of a possible history Given a phylogenetic tree T, each character X_i of the character matrix determines a function, also noted X_i, from the leaf set to the set of states $\mathcal{E}$. A possible history of X_i is a function $\hat{X}_i$ from the set of *all* nodes of the tree to $\mathcal{E}$ which coincides with X_i on the leaves. In other words, $\hat{X}_i$ assigns a state to internal nodes of T. The probability of a possible history is just the probability of the state at the root (given by the stationary distribution since the process reached its stationary state) times the probability of changes down each branch of the tree. Formally, noting T the tree, (u, v) the branch between node u and v, b_{uv} the length of branch (u, v), $\hat{X}_i(v)$ the state at node v, and Q the instantaneous rate matrix used to compute probabilities of changes, we have

$$P(\hat{X}_i|T, (b_{uv}), Q) = \pi_{\hat{X}_i(v_0)} \sum_{\text{branches (u,v)}} P_{\hat{X}_i(u)\hat{X}_i(v)}(b_{uv}) \tag{2.3}$$

where v_0 is the root of the tree.

As an illustration, for the tree and possible history drawn in Figure 2.2, we have

$$\begin{aligned} P(A, C, C, C, G, x, y, z, w|T, (t_1, \ldots, t_7), Q) = \\ P(x) \times P(y|x, t_6) \times P(A|y, t_1) \times P(C|y, t_2) \\ \times P(z|x, t_8) \times P(C|z, t_3) \\ \times P(w|z, t_7) \times P(C|w, t_4) \times P(G|w, t_5). \end{aligned}$$

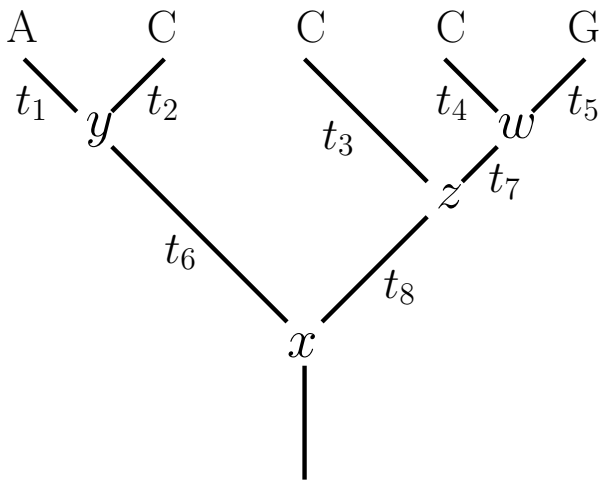

Figure 2.2: A tree (topology and branch lengths) used to illustrate likelihood computations. A,C,C,C and G correspond to characters observed in extant taxa while w, x, y, z are assigned to inner nodes to specify a possible history.

Likelihood of an observation Character states in internal nodes of the tree are unknown and we are more interested in the likelihood of an observable character that in the likelihood of a complete character. This is because we do not want the inference to be conditioned on a particular history. Instead, the probability of observing a character is a weighted average over all possible histories. This is easily done by summing equation (2.3) over all possible histories of X_i:

$$P(X_i|T,(b_{uv}),Q) = \sum_{\substack{\text{all possible} \\ \text{histories } \hat{X}_i}} P(\hat{X}_i|T,(b_{uv}),Q).$$

Going back to the example of Figure 2.2, this involves summing over all possible assignments of nucleotides at inner nodes

$$P(X_i|T,(t_1,\ldots,t_7),Q) = \sum_x \sum_y \sum_z \sum_w P(A,C,C,C,G,x,y,z,w|T,(t_1,\ldots,t_7),Q)$$

where x, y, z and t range over $\{A, C, G, T\}$.

For s leaves in the tree, this summation is over c^{s-1} possible assignments, which rapidly becomes too large to enumerate even for moderate number of species. For instance, for $s = 30$ species and DNA sequences, this is an astonishing 2.88×10^{28}.

However, the summation can be structured in a shape that mimics the topology

of T. For the example of Figure 2.2

$$P(X_i|T,(t_1,\ldots,t_7),Q) = \sum_x P(x)\Big(\sum_y P(y|x,t_6)P(A|y,t_1)P(C|y,t_2)\Big) \times\Big(\sum_z P(z|x,t_8)P(C|z,t_3) (\sum_w P(w|z,t_7)P(C|w,t_4)P(G|w,t_5))\Big) \tag{2.4}$$

where the pattern of parenthesis and terms for leaves is $(A,C)(C,(C,G))$ which is exactly the structure of the tree. The flow of computation in equation (2.4) is from the inside of innermost parenthesis outwards. It suggests a flow of information down the tree.

Pruning Algorithm Taking advantage of this remark, Felsenstein (1981) designed a pruning algorithm to evaluate the summation in an efficient way. The pruning algorithm is also known as belief propagation, a special case of sum-product algorithm, in graphical models (Pearl, 1982). It makes use of the *partial conditional likelihood* of a subtree, defined as

$$L_i^v(x) = P(X_i^v|T,(b_{uv}),Q,\hat{X}_i(v)=x),$$

where v is an internal node, x is a state and X_i^v is the restriction of X_i to descendants of node v (see Fig. 2.3). $L_i^v(x)$ is the likelihood at site i for the subtree underlying node v, conditional on state x at node v. Dropping the $T,(b_{uv}),Q$ from the notation for

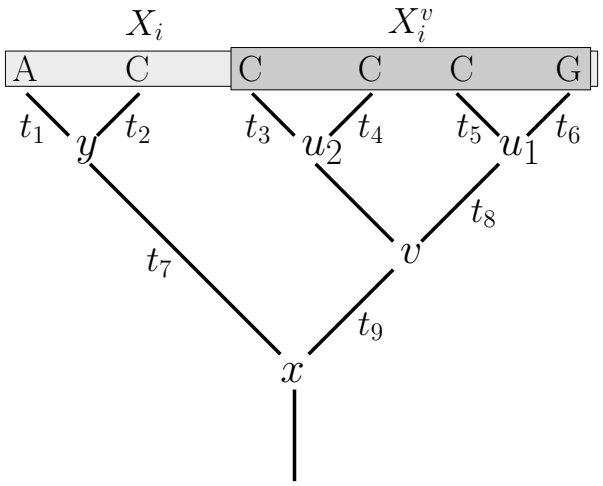

Figure 2.3: Illustration of a node v, its children u_1 and u_2, the character X_i and its restriction X_i^v to the subtree rooted at v.

convenience, the likelihood of character X_i can be expressed

$$P(X_i) = \sum_{x\in\mathcal{E}} P[\hat{X}_i(v_0)=x]L_i^{v_0}(x) \tag{2.5}$$

where v_0 is again the root node. Since we assumed the chain reached the stationary distribution, $P[\hat{X}_i(v_0) = x]$ is nothing else that π_x. The function $L_i^v(x)$ satisfies the recurrence

$$L_i^v(x) = \left(\sum_y P_{xy}(t_1)L_i^{u_1}(y)\right)\left(\sum_y P_{xy}(t_2)L_i^{u_2}(y)\right) \tag{2.6}$$

for all internal nodes v where u_1 and u_2 are the children of v and $t_1 = b_{vu_1}$ (resp. $t_2 = b_{vu_2}$) is the length of the branch connecting v to u_1 (resp. u_2). Equation 2.6 results from the assumption that evolution in different lineages is independent. We are only left with defining $L_i^v(x)$ for leaves. For leaf j we have

$$L_i^j(x) = \begin{cases} 1, & \text{if } X_{ij} = x \\ 0, & \text{otherwise} \end{cases} \tag{2.7}$$

Equation (2.7) assumes that each X_{ij} corresponds to a precise observation of state x, hence we have probability 1 to observe that state and 0 to observe any other state. In DNA sequences, if we are unable for some reason to decide between A and G at a X_{ij} we can handle the ambiguity by setting $L_i^j(A) = L_i^j(G) = 1$. The $L_i^j(x)$ do not add up to 1. This is not a problem because they correspond to different conditioning and not to different outcomes. Conditional to $X_{ij} = A$, the observation $X_{ij} \in \{A, G\}$ has probability 1 (not 1/2) hence $L_i^j(A) = 1$, and the same for $L_i^j(G)$.

Computing $L_i^v(x)$ from the leaves upward (using equation 2.6) we can compute $P(X_i)$. The transition probabilities $P_{xy}(t_1)$ and $P_{xy}(t_2)$ are determined from equation 2.2.2. This requires the diagonalization of matrix Q but we only need to do it once, after which it takes $O(c)$ operations to evaluate $L_i^v(x)$.

All calculations presented above are for a rooted. However, for a time reversible, stationary process the flow of time can run both directions. The reversibility assumption implies $\pi_x P_{xy}(t) = \pi_y P_{yx}(t)$. Substituting $\pi_y P_{yx}(t)$ to $\pi_x P_{xy}(t)$ in equation (2.5) shows that we can move the root to any children of the root without affecting the likelihood. By induction, this is also true for any node of the tree: the likelihood value is independent of the location of the root.

Complexity of the algorithm Wrapping up everything together, computing the log-likelihood involves

(i) diagonalization of Q;

(ii) computation of $e^{Q\mu t}$ for each branch of the tree, where t is the length of that branch;

(iii) computation for every possible state x, node v and site i of $L_i^v(x)$, applying equation (2.6) using a post-order traversal of the tree;

(iv) Taking the logarithm and summing over all sites.

Recall that n is the number of sites, s the number of species and c the number of states. Step (i) can be performed in $O(c^3)$ time using standard numerical techniques. Step (ii) takes $O(sc^3)$ time. Step (iii) takes $O(snc^2)$ time and step (iv) takes $O(n)$ time. The whole algorithm therefore takes $O(sc^3 + snc^2)$ time and step (iii) is by far the most computationally demanding step.

Rates across sites (RAS) and heterotachy The above calculation of the probability of observing the aligned data matrix assumes that the same set of branch lengths (b_{uv}) apply to every site in the sequence. This is another way of saying that the rate across sites are equal. This is a strong and unrealistic assumption. In real data sets, there are typically fast and slowly evolving sites. Functionally important sites are conserved during evolution while unimportant sites are free to vary.

There are two ways to relax this assumption. The first is to divide the sequence into partitions (using some exterior knowledge, such as the position in the codon for coding sequences) and attribute a different rate to each partition. Protein coding genes, for example might be divided into three partitions according to codon position.

The other way assume that the rate at a site is unknown but drawn from some distribution. Yang (1993) first introduced calculations incorporating variable rates across sites. He proposed to model the variation of rate across site by a continuous distribution: the rate of a specific site i is not a constant but a random variable $r(i)$. The likelihood of character X_i is then calculated by integrating over all possible rates:

$$P(X_i) = \int_0^\infty P(X_i|r(i) = r)f(r)dr \tag{2.8}$$

where f is the probability density function of the rate distribution and where $P(X_i|r(i) = r)$ is the likelihood of X_i, conditional on $r(i) = r$ for this site. This term is calculated by multiplying all branch lengths in the tree by r when applying the recurrence (2.6). Yang (1993) proposed a Gamma distribution for $f(r)$. Shape and scale of the Gamma distribution are $(\alpha, 1/\alpha)$ and α can be estimated from the data by maximum-likelihood method. This ensures that the mean of the Gamma is still 1 so that a branch length can still be interpreted as the expected number of substitutions on that branch. The shape parameter α is the inverse of the gamma distribution variance, high values of α correspond to (roughly) constant rates whereas low values correspond to a great heterogeneity of rates. Note that sites are not assigned with a rate in this calculation. All rates are considered and the corresponding conditional likelihoods are averaged out. Even with rates varying across sites, the sites remain independent and identically distributed.

The integration in equation (2.8) must be performed numerically, which is time consuming. In practice, Yang (1994b) discretizes the gamma distribution (usually in four classes) and assume a discrete rather than a continuous distributions across sites:

$$P(X_i) = \sum_{j=1}^{g} P(X_i|r(i) = r_j)p_j, \tag{2.9}$$

where g is the number of rate classes and p_j the probability of rate class j. The complexity of likelihood computation under the discrete-Gamma RAS model is $O(sc^3g+snc^2g)$, essentially g times the complexity of the equal rate variation. RAS models typically lead to large increase of the log-likelihood, compared to constant-rate models.

In addition to the discretized Gamma distribution, an important RAS model is the Invariant sites model. It assumes that the rate at a site is 0 with probability p and $1/(1-p)$ with probability $(1-p)$. The mean rate is again 1, branch lengths can be interpreted the usual way and the likelihood calculation is done is the same way than for Gamma distribution.

Even with rates across sites, rates are constant along a tree. A slow site is slow on every branch of the tree while a fast site is fast on every branch of the tree. However, a slow site could undergo a burst of evolution on a branch before returning to slow rates (Gu, 2006; Kitazoe et al., 2007). And even without bursts of evolution, rates might vary gradually along a branch. This idea is known as heterotachy, covarion or site-specific rate variation and has been modeled by a number of people (Galtier, 2001; Huelsenbeck, 2002; Tuffley and Steel, 1997). The covarion modeling assume Markov-modulated Markov process where the rate r of a site evolves according to a Markov process on the tree (in the simplest case between the two states "ON" and "OFF"). Several further refinement of the model exist: they assume that sites evolve under different substitution process in a more subtle way than simple RAS (Lartillot and Philippe, 2004), that the process change over time in a more subtly way than heterotachy (Blanquart and Lartillot, 2006) or even both (Blanquart and Lartillot, 2008). We do not discuss covarion models and other refinements in more details.

The model is not limited to the instantaneous rate matrix Q and the stationary distribution. RAS is also a full-fledged component of the model. RAS models where rates follow a proportion of invariant sites are noted +I, those where rates follow a discretized Gamma are noted +Γ (or +G). Finally models where some sites are invariants while the other have rates following a discretized Gamma distribution are noted +I+Γ (or +I+G). In the following, we note M for the Markov model of sequence evolution which includes at least the instantaneous rate matrix Q and can also include the options +I and +Γ. The parameters of the model include the coefficients of Q, the p proportion for +I and the shape parameter α for +Γ.

2.2.4 Optimizing the Likelihood

In this section, the topology T is constant. The problem of searching through the topology will be discussed in Section 2.3. We only discuss the problem of optimizing the parameters on a given tree.

For a given tree, one must optimize the model parameters: coefficients of Q (if M only specifies the shape of Q) and potentially p and α (again, depending on M) and the branch lengths (b_{uv}). The number of branches of an unrooted binary tree with s leaves is $2s - 3$. For a 50 species tree with a $GTR + I + G$ model, there are 108 parameters to optimize (97 branch lengths and 11 parameters for the model of sequence evolution). This is a high-dimensional, non-linear optimization problem. And there is no reason to believe the likelihood function is convex.

The peak of the likelihood landscape cannot be found analytically, except for very small trees (Yang, 2000). Instead the peak of the likelihood surface is found numerically. Branch lengths and parameters of the evolution model do not play the same role. Some of the parameters of the evolution model, such as the transition/transversion rate ration κ are difficult to estimate because it is difficult to obtain information about the slope and the curvature of the likelihood function for this parameter. It turns out that branch lengths are easier to optimize, the slope and curvature of the likelihood tell us in which direction and by how much to change the length of a branch.

Branch lengths are usually improved iteratively, one at the time. The general

approach is to

1. Choose initial branch lengths (b_{uv});
2. Repeat for each branch (uv):
   (a) Find a real number d_{uv} such that replacing b_{uv} by $b_{uv} + d_{uv}$ maximizes the likelihood (with respect to b_{uv});
   (b) Replace b_{uv} by $b_{uv} + d_{uv}$ and update the $L_i^v(x)$ values involving b_{uv}
3. If d_{uv} is small for all branches then return the current branches, otherwise go back to step 2.

Softwares differ with respect to the techniques used to determine d_{uv} but the most popular technique is Newton-Raphson. The optimization is complicated by the constraint that branch lengths must be non-negative. While the community has not gone past simple heuristics, these heuristics are proving highly effective.

2.3 Tree Search Strategies

2.3.1 Complete Search

Most inference programs take the same approach to inferring phylogenies. They first visit a tree. For that tree, they optimize the parameter such that the likelihood is maximized (Sec. 2.2.4). The maximum likelihood value for that tree and corresponding parameter values are then stored and the program moves on to another tree. The tree with largest overall maximum likelihood is the ML estimate of the phylogeny.

Exhaustive search of the tree space is just impossible. The number N_s of unrooted (unweighted) trees with s leaves increases faster than exponentially with respect to the number s of leaves. In fact, it is fairly easy to calculate N_s, which is given by the simple formula

$$N_s = (2s-5)!! = 1 \times 3 \times \ldots \times 2s-5 \propto s^s.$$

For $s = 30$ species, N_s is an overwhelming 8.68×10^{36}. And unlike the pruning algorithm for likelihood computation, there is no trick to reduce the number of visited tree to a more tractable order of magnitude while guaranteeing discovery of the ML tree. Finding the ML tree is a *NP*-hard problem.

2.3.2 Heuristic Searches: Initial Tree

We have no choice but to use heuristics to find a "reasonable" approximation to the ML tree. All heuristics discussed here start with an initial tree and then move through the tree space, accepting only moves improving the likelihood of the tree, until a local optimum is reached.

There are several methods of choosing an initial tree. The most popular are random trees, distance-based trees, addition trees and star decomposition. Random trees are most useful to explore thoroughly the likelihood landscape and make sure the search is not stuck in a local optima. Distance-based methods calculate a measure

of distance between each pair of species and then find the tree that predicts the observed distance as accurately as possible, they are a fast way to get a reasonable starting point. Addition trees randomly order the taxa, constructs a three with three of them and then sequentially adds each taxon to the tree at some place that maximizes its likelihood. When inserting the k-th taxon, we have to evaluate $2k-5$ possible trees, one for each branch of the tree with $(k-1)$ leaves constructed on up to now. In total, only $1+3+\cdots+(2s-5)$ trees are evaluated which is only a small fraction of $(2n-5)!!$. Addition trees have the nice property that different orders can give different starting trees and broaden the search of the ML tree. Star decomposition has the opposite philosophy; it starts with all species present but a completely unresolved tree (a star tree). We gradually resolve the tree by grouping two lineages at each step, until a bifurcating tree is achieved. Again, changing the groupings or the grouping method can give different starting trees.

2.3.3 Heuristic Searches: Hill Climbing

The most popular method for searching tree is local search by hill climbing. Starting from an initial tree, we perform a move on the tree. The move is a minor modification of the tree, to reach a neighboring tree. If the new tree has a higher likelihood than the initial one, we keep the new tree. We then go on and try new moves, looking for improvement of the likelihood and stop when no move can improve the likelihood. The corresponding topology corresponds to a local optimum of the likelihood in the tree space. However, nothing guarantees that it is a global optimum. The search can for example be stuck in a "peak" of the likelihood landscape, well separated from higher "peaks" by "valleys" that require several moves to be crossed. That is why thorough search of the tree space should involve multiple starting point.

For the search strategy to be efficient, the moves must satisfy a few properties. First, it should be possible to change any tree to any other in a limited number of moves. Otherwise the search would be confined in a strict subset of the space tree. Second, the moves should build upon the values of $L_i^v(x)$ to allow fast evaluation of the likelihood of a tree after a move. It should therefore preserve most subtrees of a tree. We describe 3 kind of moves with these properties: nearest-neighbor interchange (NNI), subtree pruning and regrafting (SPR), tree bisection and reconnection (TBR).

Nearest-neighbor Interchange *Nearest-neighbor interchange* (NNI) acts by swapping to adjacent branches of a tree. Another description is that NNI acts by erasing an interior branch of the tree and all branches adjacent to it, leaving out four disconnected subtrees . The four disconnected subtrees are connected again to a branch. There are in total three possible configurations: the original one and two alternatives (see Fig. 2.4). NNI examines the two alternatives. For a tree with s species, there are $s-3$ branches and thus $2(n-3)$ alternatives needs to be considered in total.

Subtree pruning and regrafting *Subtree pruning and regrafting* (SPR) acts by erasing a branch from the tree, leaving out a subtree. The subtree can then be regrafted on any branch of the remaining tree and inserts a node into the branch to which it is

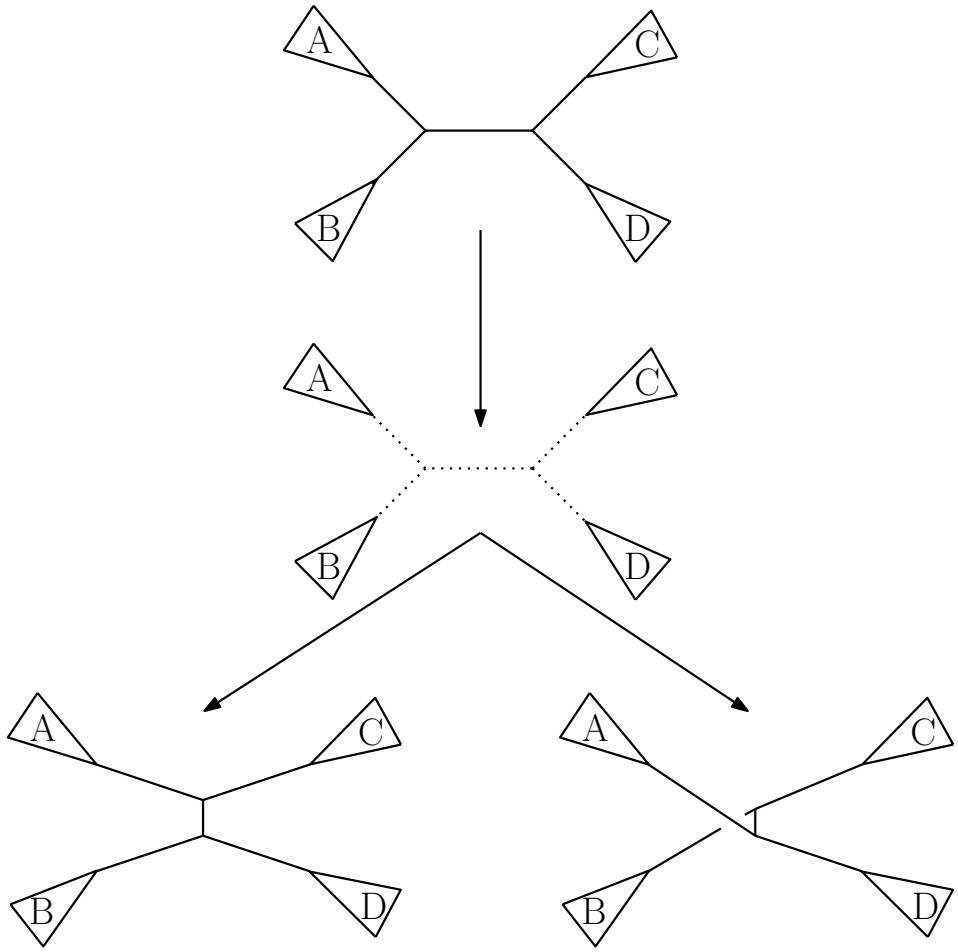

Figure 2.4: Example of nearest-neighbor interchange. An interior branch is erased and the four subtrees are reconnected in one of two possible alternatives.

attached. The move is illustrated in Figure 2.5

In a tree with s leaves, if erasing a an interior branch creates a subtree with s_1 species, the remaining tree has $s - s_1$ species and $2(s - s_1) - 3$ branches where to reconnect the subtree. One is of course the original tree so there are only $2(s - s_1) - 4$ alternatives. Considering this time the tree with $(s - s_1)$ as the subtree, there are $2s_1 - 4$ additional alternatives. Each interior branch thus generates $2s - 8$ alternatives. Each exterior branch creates $2(s - 1) - 4 = 2s - 6$ alternatives. Some alternatives may coincide but there are at most $s(2s - 6) + (s - 3)(2s - 8) = 4(s - 2)(s - 3)$ alternatives to evaluate. SPR allows for wider moves and broader search than NNI.

Tree bisection and reconnection *Tree bisection and reconnection* (TBR) acts by erasing an interior branch of the tree. The two resulting subtrees are considered as separate trees. Any branch from one tree can then be attached to any branch of the other. The move is illustrated in Figure 2.6

If removing a branch creates two subtrees with s_1 and s_2 species, TBR generates $(2s_1 - 3)(2s_2 - 3)$ trees, one of them being of course the original. Unlike NNI and SPR, there is no general formula for the number of candidates generated. It depends on

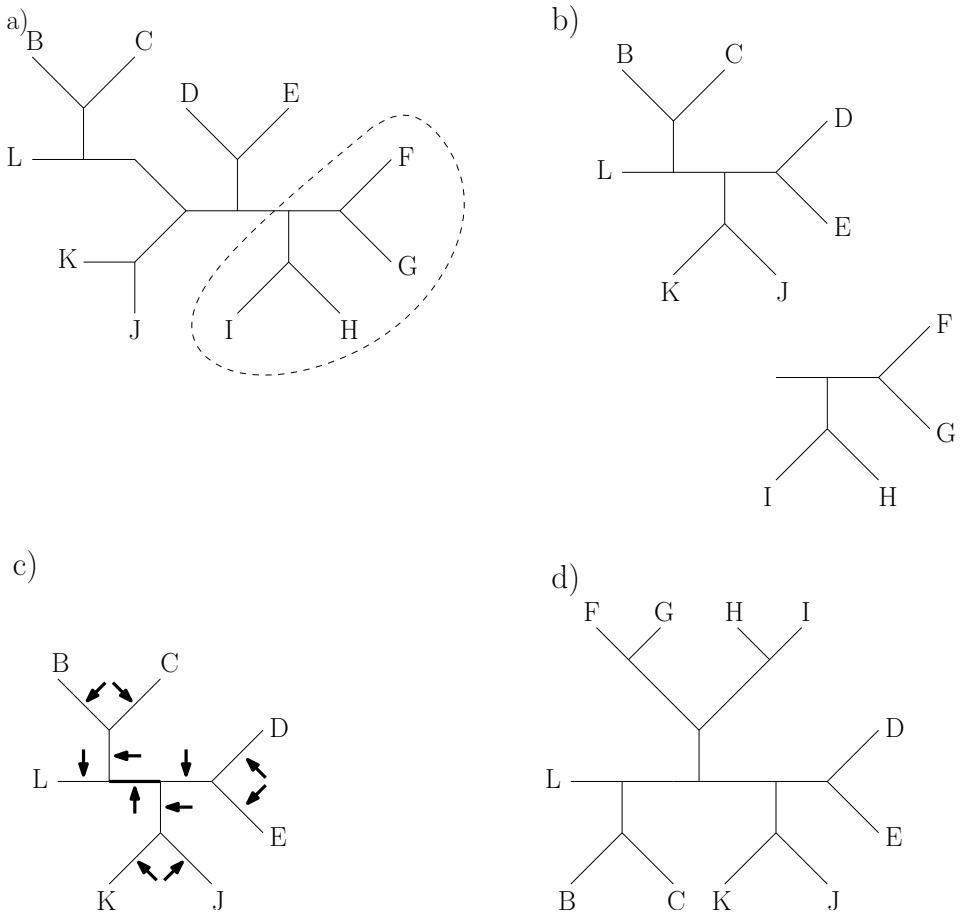

Figure 2.5: Example of Subtree pruning and regrafting. A branch of the tree is erased (a), giving the remaining tree and a subtree (b). The subtree can be regrafted on every branch of the remaining tree (c, indicated by arrows). The result corresponding to the bold branch in the remaining tree is shown (d).

the shape of the tree (Allen and Steel, 2001).

2.4 Validation and Significance of the Result

2.4.1 Phylogenetic Model

We saw throughout section 2.2 and particularly in section 2.2.1 that computing a likelihood requires a phylogenetic model (T, M) made up of several parameters of different nature: a tree T and a model of sequence evolution M. The tree T is made up of a discrete topology, also noted T, and the continuous branch lengths (b_{uv}) of this tree. The evolution model M is made of an instantaneous rate matrix Q (Sec. 2.2.2) and possibly options +I and +Γ (Sec. 2.2.3) to model the presence of heterogeneity among sites. Options +I and +Γ each require one continuous parameter ($p \in [0, 1]$ and $\alpha > 0$). When studying DNA sequences, we only specify the shape of Q and it has between 0 (JC69) and 8 (GTR) free parameters. However, when studying protein

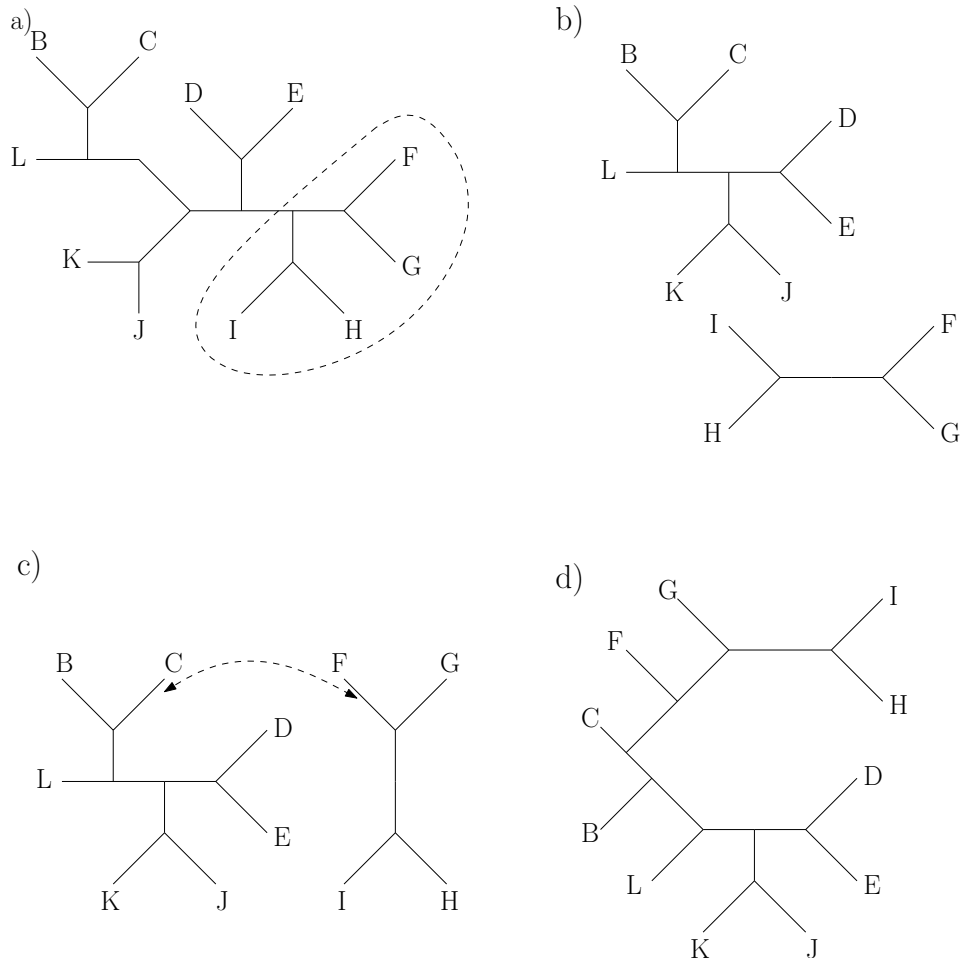

Figure 2.6: Example of tree bisection and reconnection. A branch of the tree is erased (a), giving two remaining trees (b). Any branch of one tree can be connected to any branch of the other (c, illustrated by an arrow). The result corresponding to the illustration arrow is shown (d).

sequences, Q is usually an empirical rate matrix, computed on an independent data set of closely related sequences or picked up from the literature (Jones et al., 1992; Yang et al., 1998). The specificity of protein sequences stems from the size of the alphabet, 20 against only 4 for DNA sequences, which would translate to up to 208 free parameters for Q.

When studying DNA sequences for example, we would like to test the topology, the branch lengths, p and α parameters and other parameters of the model of sequence evolution such as the transition/transversion ratio κ. However, we can not test all of them at the same time and some can not even be tested in a simple framework. In addition to testing, we would like to construct confidence intervals for the same parameters. This is also easier for some parameters than for others.

2.4.2 Hypothesis Testing

Imagine the tree T is fixed, the evolution model is set to K2P with no invariant sites nor rates across sites and we want to test $H_0 : \kappa = \kappa_0 = 1$ against $H_1 : \kappa \neq 1$. Assume furthermore that the model is correctly specified, or in other words that the sites are sampled from a K2P-Markov chain running on tree T. Standard hypothesis testing (Kendall and Stuart, 1973) suggests using log-likelihood ratio test (LRT):

$$LR = 2 \log \frac{P(X_i|T, M, \hat{\kappa})}{P(X_i|T, M, \kappa = \kappa_0)}$$

It also guarantees that LR is asymptotically a chi-square variate with 1 degree of freedom.

$$LR \sim \chi^2_1$$

The asymptotic distribution of LR can be used to build test and calculate a confidence interval of κ.

We would like to extend this result to other parameters such as the proportion p of invariant sites. With the same assumptions as before (given T and K2P model), we can test $H_0 : p = p_0 = 0$ against $H_1 : p \neq 0$. The LRT statistics is still

$$LR = 2 \log \frac{P(X_i|T, M, \hat{p})}{P(X_i|T, M, p = p_0)}$$

However, in this case the LRT is not asymptotically chi-square with 1 degree of freedom. p can take values in $[0, 1]$ and $p_0 = 0$ is at the boundary of $[0, 1]$. Standard LRT theory does not apply. Self and Liang (1987) proved that in this case

$$LR \sim \frac{1}{2}\delta_0 + \frac{1}{2}\chi^2_1$$

where δ_0 is the unit mass on 0. Since branch lengths are constrained to be positive, the same applies when testing the branch length of a given topology through $H_0 : b = b_0 = 0$ against $H_1 : b > 0$ since $b_0 = 0$ lies on the boundary of the parameter space. Ota et al. (2000) derived similar results for other continuous parameters of the phylogenetic model (T, M).

The first example is actually a test of JC69 against K2P while the second one is test of +I against no +I. But we could be interested in testing F81 against K2P+I. This is not possible through LRT because JC69 is not nested in K2P+I. Another condition for LR to be asymptotically chi-squared is that the correct model belongs to the full model. In phylogenetics, it means that the model of sequence evolution and the tree topology must be correct. This is a stringent constraint. It means that if the tree T is not given, its topology must be correctly estimated. It is essentially because phylogenetic models (T, M) with different topologies are in general not nested (Steel and Szekely, 2006b). But correctly estimating the topology is one of the inference main goal. When the correct topology is not recovered, the test may not be correctly calibrated because of the induced model bias (Buckley, 2002). Furthermore, all models are misspecified to a point because evolution is probably more complex that a nice Markov-chain running on a tree (see Sec. 5). But the biggest caveat is perhaps that likelihood ratio testing

procedures apply to real valued estimates whereas tree topology (or even the tree itself) is not real valued.

To sum up, for a given or estimated tree, LR testing can be used to test nested models against each other (*e.g.* JC69 against K2P, K2P against GTR+I+Γ,...) or to test interior branch lengths. If the models is misspecified, results should be interpreted with caution. Finally, LR testing can not be used in general to test topologies, with the exception of a multifurcating tree against a fully resolved tree.

2.4.3 Testing Topologies

Instead of using LR testing, test of topologies reduce a tree to its log-likelihood score and compare the difference in log-likelihoods between trees with different topologies (see Chap. A). As sequences grow infinitely long, the likelihood ℓ_n^T of a tree T converges to its asymptotic value ℓ^T. The best of two trees T and T' is the one with higher likelihood. If two trees have the same likelihood ($\ell^T = \ell^{T'}$), then the difference in log-likelihood at each site is drawn from some distribution with expectation 0.

The most popular likelihood-based tests of topology are based on this simple premise (Goldman et al., 2000). If Z is the difference between T and T' in log-likelihood at a site, they test $H_0 : E[Z] = 0$ against $H_1 : E[Z] \neq 0$. The simplest one is the Kishino-Hasegawa (KH) test (Hasegawa and Kishino, 1989; Kishino and Hasegawa, 1989). It uses bootstrap sampling to infer the distribution of the sum of differences and test whether $E[Z]$ is significantly different from 0. Since one is only interested in testing topologies but nevertheless has to compute likelihood values for each site, there are many nuisances parameters: branch lengths, parameters of the evolutionary model and potentially the evolutionary model M itself. The evolutionary model M is often fixed. Branch lengths and parameters of the evolutionary model can be optimized for each bootstrap data set or fixed once and for all to their estimate on the original data set. The most popular method is resampling estimated log-likelihood (RELL); nuisance parameters are optimized once and for all on the original data set. With the RELL method, the difference Z_i in log-likelihood for site i needs to be computed only once and we just have to resample the Z_i values to see whether 0 lies in the tail of the distribution. RELL is a time-saving approximation that requires the evolutionary model M to be correctly specified and the amount of data to be large enough for the approximation to be valid but still has been shown to perform well (Hasegawa and Kishino, 1994).

There are at least two caveats with the KH test. The first, noted by Swofford et al. (1996) and fully explained by Shimodaira and Hasegawa (1999) is the topology choice. T and T' are assumed to be specified independently of any analysis of the data used for the testing. However, the KH test is often used to compare the maximum-likelihood tree $T = T_{ML}$ to an a priori tree or to the next best tree T'. In this case, $E[Z] > 0$ and the test calibrated under the null hypothesis $H_0 : E[Z] = 0$ is no longer valid anymore. Another caveat is the way KH test is used to construct confidence set for trees; many trees are tested again the best tree and all trees not rejected by the KH test are included in the confidence way. This is not the proper way of doing multiple tests and accept too few trees (Goldman et al., 2000). Shimodaira and Hasegawa

(1999) introduces the Shimodaira-Hasegawa (SH) test that addresses these two issues. Multiple topologies $T_1, \ldots, T_N$ can be assessed and the SH tests $H_0 : \ell^{T_1} = \ldots = \ell^{T_N}$ against H_1 : "Not all ℓ^{T_i} are equals". By assuming all trees have equal likelihoods, SH works under a "worst case" assumption. It is conservative and quickly turns unable to reject any tree when the number of trees to test grows too large.

Other multiple comparison tests include Bar-Hen and Kishino (2000) and the SOWH test (Swofford et al., 1996). Bar-Hen and Kishino (2000) compute the covariance matrix of the log-likelihood per site for several trees and use the asymptotic normality of the log-likelihood per site to build a confidence set on the topologies. SOWH test is similar to SH test but generates sites under a well-chosen phylogenetic model (T, M) instead of resampling them from the original data set.

2.5 Thesis Outline

We briefly return to the questions raised at the beginning of section 2.2. The model for sequence evolution was described in section 2.2.2, an efficient way to compute the likelihood was presented in section 2.2.3 and optimization of the parameter was discussed in section 2.2.4 for continuous parameters and section 2.3 for the topology. The last question about validation and significance of the result yet is briefly discussed in section 2.4 through test of topologies.

We do not address this question any longer in the introduction. It is discussed at length in the rest of this thesis. We only sketch the issues studied in part I and part II. Each part starts with a specific introduction, which presents the background of the research and motivates it. It continues with the main results, presented as research articles, accepted or submitted to peer-reviewed journals. It ends with a discussion of the results and prospects for further research.

2.5.1 Stochastic Errors

Sampling process In Part I, we are interested in errors induced by the stochastic nature of the observations. The first kind of error is induced by the sampling of sites. For a given model M of sequence evolution (with parameters), we can associate to each tree T (with branch lengths) a quantity ℓ^T which represents the asymptotic likelihood of that tree under model M. The ℓ^T values are used to rank the trees from best to worst. Under model M, the maximum likelihood tree is in fact exactly the tree maximizing ℓ^T. However we cannot calculate ℓ^T for any tree. It would require perfect knowledge about the sequence evolution process, hereafter noted Q and of which M is but an approximation, or infinitely many observations. Both are out of our reach. For want of anything better, we usually replace ℓ^T with ℓ^T_n, the likelihood of T for a sample of size n from Q. Although ℓ^T_n converges to ℓ^T as n grows to infinity, the sampling process makes it fluctuate around ℓ^T. Moreover, such fluctuations can invert the ranking of trees T and T', so that $\ell^T_n < \ell^{T'}_n$ while $\ell^{T'} < \ell^T$. With n observations only, the maximum-likelihood method has a positive chance to prefer T' to T when it ought to choose T based on their asymptotic likelihood. In Chapter A, we use

concentration techniques to upper bound the fluctuations of ℓ_n^T around ℓ^T. Using the same techniques, we also bound the probability of preferring a tree to another tree with higher likelihood because of observations sampling.

Change in the generating process A second kind of error arise from inconsistency of the generating process. All sites are generated by the sequence evolution process Q and maximum likelihood selects, for a given model of sequence evolution M, the tree T that is closest to Q, measured by the Kullback-Leibler divergence between Q and $P(.;T,M)$. As we increase sequence lengths, the likelihood score ℓ_n^T converges to ℓ^T. It is thus tempting to use sequences as long as possible to estimate ℓ^T through ℓ_n^T. However, if the generating process changes to Q' at some point along the sequence, so does ℓ^T. Far for improving the accuracy, using observations sampled from Q' to calculate ℓ_n^T will actually hinder the estimation of ℓ^T and ultimately of T. Before including new observations to the analysis, it is essential to assess that they are consistent with previous observations. New observations can be sites coming from a newly sequenced gene. Changes in the generating process Q' are a potential source of error only if they shift ℓ^T. Estimating ℓ^T is equivalent to estimating the mean of some distribution. Using phylogenetics as a motivating example, we develop in Chapter B a non-parametric test to test shifts in the mean of the generating distribution of sequential data. The intuitive idea is that shifts of the mean induced by adding observations to the analysis should be equivalent as shifts of the mean induced by removing observations. We make this intuition correct with Edgeworth expansions. Edgeworth expansions also give the first order correction term in the approximation, valid both for continuous and discrete observations.

2.5.2 Detecting Outliers

Robust phylogenies and influent sites In Part II, we are interested in errors induced by outliers, also referred to as aberrant points. We saw in section 2.1.3 that an inference method should be robust to small violations of the model. It should also be resistant to outliers. Indeed, if the method is overly sensitive to the data, changing as few as a handful of sites can dramatically modify the inferred tree. The significance of such a tree should be seriously questioned. There are several ways to construct outlier resistant methods. The most straightforward is of course to remove outliers from the analysis, or at least to weight them down. But to do so requires some prior knowledge about the outliers. The first step is thus of course to identify outliers. Using tools coming from the robustness analysis such as influence function and sensitivity curve, we present in Chapter C an adaptation of influence function to the detection of influent sites. This index is applied to a real-case study of a fungal phylogeny and successfully identifies peculiar sites, with high influence on the phylogeny estimate.

Influent taxon Sites in the alignment play an important role in providing the information necessary to unravel the (evolutionary) structure of the taxa included in the data matrix. However, taxa also play a full fledged part in the inference and adequate taxon sampling is all but superfluous. Small changes in the taxon sampling can have dramatic changes on the topology. Like with sites, a tree overly sensitive

to taxon sampling should be seriously questioned. Again, constructing phylogenies resistant to rogue species requires preliminary detection of influent species. A natural way to measure the influence of a species is via jackknifing of species. Unlike sites, jackknifing of species is a peculiar statistical problem as species are not independent. Nevertheless, we present in Chapter D an adaptation of influence function to the detection of influent species. This index is applied to a real-case study of mammal phylogeny and successfully identifies taxa previously identified in the literature as rogue taxa, with hindering effect on the inference process.

Part I

Stochastic Errors

Summary

In this part, we use concentration tools and Edgeworth expansions to control two sources of variability: site sampling and inconsistency of the sampling distribution along the observations. Site sampling and sudden changes in the sampling distribution are major concerns in any estimation problem, including but not limited to phylogenetic inference. We start Chapter 3 with an overview of concentration inequalities and a description of Edgeworth expansions before presenting the sources of variability and their significance in the specific context of phylogenetics.

The expected likelihood score of a tree represents the extent to which it explains the data. Because of the finite number of observations, we can only estimate it and the estimation can be quite far from the exact value. In Chapter A, we use concentration inequalities to upper-bound, for a given tree, the fluctuations of the likelihood score around its expected value. Using the same tools, we also bound the probability of choosing a suboptimal tree from a given set because of the limited number of observations.

Inconsistency of the sampling distribution introduces a bias in the inference process by changing the score of a tree. The score can be represented as the mean of a transformation of the sampling distribution. In Chapter B, we develop a procedure based on resampling techniques to test and detect changes in the mean of the distribution of sequential data. The procedure relies on Edgeworth expansions and is valid even for discrete variables, even though Cramér's regularity condition does not hold in this case. Observations labeled as drawn from another sampling distribution can then be withdrawn from the analysis in order to reduce the bias.

We conclude in Chapter 4, by discussing our results on bounding the variability. We discuss generalization of the results, caveats calling for attention and potential ways to solve them. We also discuss possible subjects for further research.

Chapter 3

Introduction

3.1 Concentration Inequalities for Sums of Independent Random Variables

Let X be a real-valued random variable. The value of X is known through its distribution function F_X. When X is a simple random variable, F_X is also simple and the behavior of X is easy to control. For example, if X is a Bernoulli variable with parameter p, $F_X = (1-p)H(x) + pH(x-1)$ where H is the Heavyside function defined as:

$$\begin{cases} \mathrm{H(x)} &= 0 \text{ if } \mathrm{x} < 0 \\ \mathrm{H(x)} &= 1 \text{ else.} \end{cases}$$

And the behavior of X is known from a glance at F_X.

For more complex X, F_X rapidly becomes too complex to be informative anymore. In the simple case where X takes value 1 and -1, each with probability $1/2$, the mean $S_n = \frac{X_1+\cdots+X_n}{n}$ of n independent copies of X has exactly $n+1$ atoms. However most of these atoms have a very small mass and correspond to rare events.

The Law of Large Numbers (LLN) states that $S_n \xrightarrow[n\to\infty]{a.s.} 0$. Even if S_n can take $n+1$ values, only those close to 0 are relevant: for large enough n, S_n is close to its expectation $E[X] = 0$ with high probability. S_n is a textbook example of random variables concentrated around their means. Concentration inequalities specify what "close" and "large probability" mean.

Our purpose in this part is to derive concentration inequalities for evolutionary trees, but before proceeding to evolutionary trees, we briefly recall inequalities bounding tail probabilities of sums of independent random variables.

First of all we recall the essential basic tools used to prove concentration inequalities. For any nonnegative random variable X,

$$E[X] = \int_0^\infty P(X \geq x)dx.$$

This implies *Markov's inequality:* for any nonnegative random variable X and $t > 0$,

$$P(X \geq t) \leq \frac{E[X]}{t}.$$

It follows from Markov's inequality that for any strictly increasing nonnegative-valued function ϕ and for any random variable X and any real number t,

$$P(X \geq t) = P(\phi(X) \geq \phi(t)) \leq \frac{E[\phi(X)]}{\phi(t)}.$$

Taking $\phi(x) = x^2$ in the previous inequality results in *Chebyshev's inequality:* for any random variable X and $t > 0$,

$$P(|X - E[X]| \geq t) = P(|X - E[X]|^2 \geq t^2) \leq \frac{E[|E - E[X]|^2]}{t^2} = \frac{Var(X)}{t^2}.$$

More generally taking $\phi(x) = x^q \quad (x \geq 0)$, for any $q > 0$ we have

$$P(|X - E[X]| \geq t) \leq \frac{E[|E - E[X]|^q]}{t^q}.$$

For a specific random variable X and deviation t, we can choose the value of q to optimize the obtained upper bound. Such moment bounds often provide very sharp estimates of the tail probabilities. A related idea is at the basis of *Chernoff's bounds*. Taking $\phi(x) = e^{sx}$ where s is an arbitrary positive number, for any random variable X and any $t > 0$, we have

$$P(X \geq t) = P(e^{sX} \geq e^{st}) \leq \frac{E[e^{sX}]}{e^{st}}.$$

And we then find an $s > 0$ that minimizes the upper bound.

Next we recall the essential inequalities for sums of independent random variables. The main concern is bounding probabilities of deviations from the mean, that it is bound $P(S_n - E[S_n] \geq t)$ where $S_n = \sum_{i=1}^n X_i$ and $X_1, \ldots, X_n$ are independent real-valued random variables.

Chebyshev's inequality and independence immediately imply

$$P(|S_n - E[S_n]| \geq t) \leq \frac{Var(S_n)}{t^2} = \frac{\sum_{i=1}^n Var(X_i)}{t^2}.$$

Rescaling the observations and writing $\sigma^2 = \frac{1}{n}\sum_{i=1}^n Var(X_i)$,

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i - E[X_i])\right| \geq t\right) \leq \frac{\sigma^2}{nt^2}.$$

Chernoff's bounds are especially convenient for bounding tail probabilities of sums of independent random variables. The reason is that since the expected value of a product of independent random variables is the product of the expected values, Chernoff's bounds have the simple form

$$P(S_n - E[S_n] \geq t) \leq e^{-st} E\left[exp\left(s\sum_{i=1}^n (X_i - E[X_i])\right)\right] = e^{-st}\prod_{i=1}^n E\left[e^{s(X_i - E[X_i])}\right].$$

The problem of finding tight bounds boils down to finding a good upper bound of the moment generating function of the random variables $X_i - E[X_i]$. There are many ways of doing this. For bounded variables, the most elegant version is perhaps due to Hoeffding (Hoeffding, 1963).

Lemma 1 (Hoeffding's Inequality) *Let X be a random variable with $E[X] = 0$ and $a \leq X \leq b$ almost surely. Then for $s > 0$,*

$$E[e^{sX}] \leq e^{s^2(b-a)^2/8}.$$

This lemma immediately implies Hoeffding's tail inequality (Hoeffding, 1963):

Theorem 2 (Hoeffding's Inequality) *Let $X_1, \ldots, X_n$ be independent real-valued random variables such that $a_i \leq X \leq b_i$ almost surely. Then for any $t > 0$ we have*

$$P(S_n - E[S_n] \geq t) \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}$$

and

$$P(S_n - E[S_n] \leq -t) \leq e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2}.$$

One caveat of Hoeffding's inequality is that it ignores information about the variance of the X_i. Bennett's and Bernstein's inequality (Bennett, 1962) provide an improvement in this respect.

Theorem 3 (Bennett's Inequality) *Let $X_1, \ldots, X_n$ be independent real-valued random variables with zero mean and assume that $|X_i| \leq M$ almost surely. Let*

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n Var(X_i).$$

Then for any $t > 0$,

$$P\left(\frac{1}{n} \sum_{i=1}^n X_i > t \right) \leq \exp\left(-\frac{n\sigma^2}{M^2} h\left(\frac{Mt}{\sigma^2} \right) \right)$$

where $h(u) = (1+u)\log(1+u) - u$ for $u \geq 0$.

Theorem 4 (Bernstein's Inequality) *Let $X_1, \ldots, X_n$ be independent real-valued random variables with zero mean and assume that $|X_i| \leq M$ almost surely. Let*

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n Var(X_i).$$

Then for any $t > 0$,

$$P\left(\frac{1}{n} \sum_{i=1}^n X_i > t \right) \leq \exp\left(-\frac{nt^2/2}{\sigma^2 + Mt/3} \right)$$

For σ^2 small compared to t, the upper bound behaves like $e^{-3nt/2M}$ instead of the slower $e^{-nt^2/2M^2}$ guaranteed by Hoeffding's inequality.

3.2 Evolutionary Trees

Hoeffding's, Bennett's and Bernstein's inequalities hold for *real-valued* sums of independent random variables. Unfortunately, evolutionary trees are more complex: as described in Section 2.2.1, evolutionary trees with s leaves are made of a discrete topology and continuous branch lengths and can be thought of as points of a large dimension manifold of $\mathbb{R}^{2^s}$ (Billera et al., 2001). Furthermore, although several metrics on the tree space exist, none is unanimously regarded as better than the other. Instead of focusing directly on evolutionary trees, we replace them by their likelihood score, which is a real-valued random variable.

Formally, consider a discrete space $\mathcal{A}$ and $X_1, \dots, X_n$ a sequence of $\mathcal{A}$-valued independent identically distributed (i.i.d) random variables with distribution Q. Furthermore consider a discrete set $\mathcal{M}$ of models such that for all $m \in \mathcal{M}$, there exists a probability distribution P_m over $\mathcal{A}$. Define the *best model* $m^\star$ as

$$m^\star = \arg\min_{m \in \mathcal{M}} \left\{ KL(Q, P_m) = \sum_{a \in A} Q(a) \ln \frac{Q(a)}{P_m(a)} \right\}$$

or any $m^\star$ if there are multiple m in the argmin. $m^\star$ is the model which induces the distribution closest to Q, in the sense of the Kullback-Leibler divergence. But Q is usually unknown, and we only have access to $m_n^\star$, the model which induces the distribution closest to Q_n, the empirical version of Q defined by $Q_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ where δ_x is the Dirac delta function.

For each model m, $Z_i^{(m)} = \log(P_m(X_i))$ is then a real-valued random variable. Incidentally, $m^\star$ is also the model m maximizing $E[Z^{(m)}]$. But again, $E[Z^{(m)}]$ is usually replaced by $E_n[Z^{(m)}] = \frac{1}{n} \sum_{i=1}^n Z_i^{(m)}$ for lack of a better alternative.

Finding the best model is thus equivalent to maximizing the expected value of the transformation $Z_i^{(m)}$ of the original variables X_i. Since the ordering of the $E[Z^{(m)}]$ gives the best model, it is essential to ensure that the $E_n[Z^{(m)}]$ are tightly concentrated around their expected value $E[Z^{(m)}]$.

In a phylogenetic study of s species, $\mathcal{A}$ is the set of s-uple of nucleotides $\mathcal{A} = \{A, C, G, T\}^s$, a model $m = (T, M)$ is made of a topology T topology with s leaves and a Markovian evolution model M (see sec 2.2). $\mathcal{M}$ is just the set in which T and M can vary. M is usually chosen from one of the models detailed in section 2.2.2 and fixed before the inference whereas T ranges in all binary topology. Each model $m = (T, M)$ induces a probability distribution over $\mathcal{A}$ as was seen in section 2.2.

Chapter A details the specificity of phylogenetic inference and how they relate to concentration theory. It also provides concentration inequalities to upper bound the tail probabilities of two family of events. The first kind of event tell us that $E_n[Z^{(m)}]$ is close to its expected value $E[Z^{(m)}]$ with high probability, with quantitative evaluation of "close" and "high probability" depending on m. The second kind of event tell us that $E_n[Z^{(m)}] - E_n[Z^{(m')}]$ and $E[Z^{(m)}] - E[Z^{(m')}]$ have the same sign with high probability with quantitative evaluation of "high probability" dependent on m and m'. The second result is highly significant as it quantifies how often we pick the worst of two models when only a sample of size n is available. Note that the quality

$E[Z^{(m)}]$ of a model m is highly dependent on both the topology T used to summarize evolutionary history and the Markovian evolution model M used for DNA sequences. For a given topology T, the value $E[Z^{(T,M)}]$ changes with the markovian model M so that the best topology for an evolution model M may not be the best for a different evolution model M'. Comparing model m and m' different only with respect to their topology is thus meaningful only if the markovian model M is appropriate.

3.3 Consistency of the Generating Process Along the Sequence

Concentration inequalities for upper bounding tail probabilities of deviations from the mean as discussed in Chapter A are exact and hold for any n no matter how small it is, unlike bounds obtained from gaussian or Poisson approximation which hold only for large n. But as a result, the upper bounds thus obtained are usually looser than bounds obtained from a gaussian/Poisson approximation.

Bernoulli variables provides perhaps the simplest illustration of such a phenomena. Let $X_1, \ldots, X_n$ independent Bernoulli random variables, all with parameter p/n. Then for any $\varepsilon > 0$, by Hoeffding's inequality

$$P\left(\sum_{i=1}^{n} X_i - p \geq n\varepsilon\right) \leq e^{-2n\varepsilon^2}$$

Whereas a Poisson approximation would approximate $\sum_{i=1}^{n} X_i$ by a Poisson random variable with parameter p: $P(p)$ which has tail probability of order $e^{-n\varepsilon}$. For $\varepsilon < 1$, the Poisson approximation is much sharper that the concentration inequality, although it holds only for large enough n. Concentration inequalities developed in chapter A require stringent conditions on the unknown generating distribution Q of the observations.

When dealing with DNA sequences, it is increasingly common that more than one gene is available. However strong the temptation to use all the genes simultaneously in the analysis, one needs to be cautious. Indeed, thanks to recombination, selective sweep, selection or other biological process, the genes may have very different evolutionary histories and adding a gene to the analysis may contaminate the sample and pollute the analysis. It is thus essential to assess whether all the genes have the same evolutionary history, or, when the genes come in sequential order, whether the last available gene is comparable to and shares a common phylogenetic signal with the previous ones.

Formally, we want to test whether all the data in the sample come from the same generating process. We develop in Chapter B a test to address this issue. The test is based on a simple statistic and makes as few assumptions as possible on the generating process. The rationale of the statistics is that under the null hypothesis of consistency, adding a few observations to the analysis should have the same effect on the estimated mean as removing a few observations from the analysis. Hence, we can approximate the distribution of shifts of the estimated mean by resampling techniques and see

whether the actual value lies in the tail of the distribution. Edgeworth expansions give the first-order correcting term for the approximation.

Appendix A

Concentration Inequality for Evolutionary Trees

This section is a modified version of the article *Concentration inequality for evolutionary trees* accepted for publication in Journal of Multivariate Analysis.

M. Mariadassou, A. Bar-Hen, Concentration inequality for evolutionary trees, *Journal of Multivariate Analysis* (2009), in press, doi:10.1016/j.jmva.2009.02.015

Abstract Maximum likelihood inferred topologies are commonly used to draw conclusions in evolutionary biology and molecular evolution. Considering the sampling error when estimating the topology is a critical issue. Bootstrap-based methods are the most popular tools to assess the robustness of clades, *i.e.* the stability of a tree and sub-trees. Unfortunately, there is no analytical result to connect the bootstrap values to the sampling variability, or at least to the number of sites and species in the study. Using concentration measure tools, we first bound the variations of the computed likelihood around its true value and then bound the sampling variability of likelihood as measured by bootstrap. In particular and unlike most bootstrap-based methods, these bounds are explicitly sensitive to both the number of species and of nucleotides.

Keywords: Bootstrap, Phylogeny, Robustness, Concentration Inequality

A.1 Introduction

Phylogenies, or evolutionary trees, are the basic structures necessary to analyze differences between species. Several methods are available to infer phylogenies, the two most popular being Maximum Parsimony (MP) and Maximum Likelihood (ML) (see Felsenstein (2004); Gascuel (2005) for a comprehensive review). The MP method has a lower computational burden when inferring phylogenies but the ML method (Buschbom and von Haeseler, 2004; Felsentstein, 1983) provides a statistical framework to the inference problem. In this chapter, we focus on ML methods and the stability of the inferred phylogeny.

A common problem is the support given to a clade, *i.e.* a subtree of particular interest. Several bootstrap methods have been developed to address specifically this issue (see (Efron et al., 1996; Felsenstein, 1985; Felsenstein and Kishino, 1993; Penny and Hendy, 1986; Shimodaira, 2002) and (Holmes, 2003) for a review). Stability is a fundamental property for a phylogeny: after inferring a tree, we want to draw some conclusions from it. For example if a phylogeny positions species A and B in the clade, it is important to assess the significance of the classification: is the clade supported by a lot of evidence or is it here "just by chance" ? Therefore, the tree must be as robust as possible: a small modification in the data should not drastically change the phylogeny and invalidate the conclusions, or at least if it does, it should only do so with a small probability. An inferred phylogeny not satisfying this property is of little use: no biological conclusions drawn from it would be reliable.

Most bootstrap methods are based on re-sampling with replacement (Efron, 1979): they mimic the true distribution of the data by the one corresponding to the sample. Doing so, they replace the true variability with the observed one whereas it can be quite different: conclusions are very dependent on a specific sample. Bootstrap methods also discard the relation between the size of the data, the number of species in the study and the stability of the phylogeny (Efron et al., 1996; Hillis and Bull, 1993; Newton, 1996).

In this paper we propose an analytical alternative to bootstrap for comparing two phylogenies. Rather than working on the phylogeny, we work on its likelihood

score: a stable ML phylogeny is equivalent to stable scores and thus to a stable score ranking. Using measure concentration tools, we obtain bounds on the probability that the empirical likelihood wanders too far away from its expectation. We also bound the probability than a given phylogeny is erroneously scored better than another one "just by chance". An advantage of doing so is to reduce the study of phylogenies and phylogenetic trees to the much simpler study of likelihood scores taking values in $\mathbb{R}$.

Section A.2 is devoted to the framework. We also introduce the notations and the main concepts. Then, in Section A.3, we derive our main result and apply it to the stability of phylogenies. Finally in Section A.4, we illustrate the method on an example, compare it to other popular methods and discuss the pros and cons. Technical proofs of some results are postponed in the appendix.

A.2 Framework

We introduce in this section the statistical framework and the notations.

A.2.1 Notations and definitions

We consider a $s \times n$ matrix $\mathcal{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n) = (X_{ij})_{i=1\ldots n, j=1\ldots s}$ representing a set of molecular sequences aligned over different species. n is the length of the sequence after the alignment (including gaps) and s the number of species. X_{ij} takes value in an alphabet $\mathcal{A}$ and codes for the state of the ith nucleotide of the alignment in the jth species. The jth line of $\mathcal{X}$ then represents the aligned sequence of species j.

When working with DNA sequences, $\mathcal{A}$ is usually a four-letter alphabet $\{A, C, G, T\}$ but it can take others values, for example when working with protein sequences (20 possibles amino-acids). The statistical unit of interest is the column $\mathbf{X}_i$, a s-dimensional vector valued in $\mathcal{A}^s$, which codes for the pattern of nucleotide i over all s species.

We assume that the pattern $\mathbf{X}_i$s are *i.i.d* random variables whose common discrete probability is Q. Although the independent sites assumption is unrealistic, it is a reasonable working hypothesis for many reasons. First, very few models account for neighbor-dependent nucleotide substitution process (Bérard et al., 2008). Second, all models used in molecular phylogenetics suppose independent sites (and indeed no dependent sites model is implemented in the most popular phylogenetic packages such as PAUP* (Swofford, 2003) or PHYLIP (Felsenstein, 2005)). Finally, apart from some extreme cases, the dependence is modeled with strong mixing conditions and its main effect can be thought as reducing the effective sampling size.

Various authors extended Hoeffding-type inequality to dependent variables cases. The core of the extension is the definition of the dependence among the variables. These bounded probabilities are exponentially bounded but the decay is related to the kind of dependence (Van de Geer, 2002). One may notice that some extension are perfectly adapted to the case of phylogeny (Tang and Yongqiang, 2007) but the statistical properties of the model or the algorithmic part necessary to compute the likelihood is not yet developed.

Definition 5 *A phylogenetic model $m = (T, M)$ is defined as the union of:*
(i) the evolution model*: the markovian evolution model M and associated parameters,*
(ii) the tree*: the topology T and associated branch lengths $\mathbf{b}_T$.*

A phylogenetic model m is basically the probabilistic model used to describe the changes between nucleotides for a given set of species and compute the likelihood of any given pattern. Although the main interest usually lies in the tree, or even only in the topology, the markovian model M is essential in computing the likelihood function and need to be chosen carefully.

Several evolution models have been proposed ranging from the simple *Jukes-Cantor* (Jukes and Cantor, 1969) to the *General Time Reversible (GTR)* (Lanave et al., 1984) including *Kimura two-parameters (K2P)* (Kimura, 1980) (see Nielsen (2004) or section 2.2.2 for more about DNA evolution models), all of them boiling down to continuous time reversible time Markov chain with more or less sophisticated rate matrix. Q is the *true* pattern distribution and, as reality is often more complex than the model used to describe it, has no reason to coincide with a Markov-chain ran along a tree (for example correlated evolution could occur on different parts of the tree). The main goal of phylogenetic inference is the to retrieve, among all distributions obtained from a phylogenetic model, the one closest to Q for the *Kullback-Leibler distance* (KL -distance) (Kullback and Leibler, 1951).

Calculating the log-likelihood $\log \mathbb{P}(x; m)$ of an observed pattern x under model m is the cornerstone of ML analysis but is in general quite difficult. Fortunately, for any Markovian evolution model, Felsenstein's pruning algorithm (Felsentstein, 1983) makes it possible (see section 2.2.3 for details).

Definition 6 *The* empirical *(resp.* true*) mean log-likelihood ℓ_n^m (resp. ℓ^m) is the mean of* $\log \mathbb{P}(\mathbf{X}; m)$ *under the empirical (resp. true) distribution:*

$$\ell_n^m = \mathbb{E}_{Q_n}[\log \mathbb{P}(\mathbf{X}; m)] = \frac{1}{n} \sum_i \log \mathbb{P}(\mathbf{X}_i; m) \tag{A.1}$$

$$\ell^m = \mathbb{E}_{Q}[\log \mathbb{P}(\mathbf{X}; m)] = \sum_{x \in \mathcal{A}^s} Q(x) \log \mathbb{P}(x; m) \tag{A.2}$$

with the empirical distribution *of patterns defined as:*

$$Q_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\mathbf{X}_i}$$

The empirical distribution is opposed to the unknown *true distribution* Q of the patterns, which is unachievable, as it would require infinite length sequences (see Kim (2000) for a geometrical interpretation). Although we should work with the ℓ^m, only the ℓ_n^m are available, which induces some stochastic fluctuations in the inference process. In the following, m is fixed and we study how good the approximation of ℓ^m by ℓ_n^m is.

We need the expressions of both the empirical and true log-likelihood. The goal, as presented in more details in Sec. A.3 is to compare not only the empirical mean log-likelihood to its true average, *i.e.* $\mathbb{P}(|\ell_n^m - \ell^m| > \varepsilon)$, but also the rankings induced on phylogenetic models by both the empirical and the true mean log-likelihood, *i.e.* the probability of $\{\ell_n^m < \ell_n^{m'}\}$ knowing $\ell^m > \ell^{m'}$.

A.2.2 Connection between ℓ^m and Q

We start by defining some vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_n$ coding for the same information as the distributions Q and Q_n. We do so because vectors are easier to manipulate than distributions.

Definition 7 *$\mathcal{N}_s$ is the support of Q,* i.e. *the subset of $\mathcal{A}^s$ made of all the patterns x with $Q(x) > 0$.*

Definition 8 *$\boldsymbol{\theta} = (\theta^x)_{x\in\mathcal{N}_s}$ (resp. $\boldsymbol{\theta}_n = (\theta_n^x)_{x\in\mathcal{N}_s}$) is the probability vector corresponding to Q (resp. Q_n),* i.e. *the vector of length $|\mathcal{N}_s|$ such that for all $x \in \mathcal{N}_s$ $\theta^x = \mathbb{P}_Q(\mathbf{X} = x)$ (resp. $\theta_n^x = \mathbb{P}_{Q_n}(\mathbf{X} = x)$).*

Note than $|\mathcal{N}_s|$ can be as large as the entire space $|\mathcal{A}^s|$ when the s species at hand are very different and as small as 4 when they are very close. When Q is a Markovian-process over a tree, $\mathcal{N}_s$ is exactly $\mathcal{A}^s$ (except for extreme cases, such as terminal branches of length 0). However, $\mathcal{N}_s$ is several orders of magnitude more restricted than $\mathcal{A}^s$ for several reasons: irreversible changes on a branch, convergent evolution in different parts of the trees, purifying selection in some branches, etc. Another argument comes from practical considerations on the alignment matrix $\mathcal{X}$. For alignment to be possible, the sequences must be fairly well conserved: no alignment is deemed reliable when the sequences are too divergent. As a result, the alignment matrices used in phylogenetic inference often have low diversity and changes in a site are sparse. Typically a significant fraction of the sites are invariant (i.e. of the type *xxxxxxx* where x is a nucleotide) and even variable patterns have one (and less frequently two) nucleotide shared among most species, and another one shared by the remaining species (i.e. up to a reordering of the species, are of the type *xxxxyyyy* where x and y are two different nucleotides). Extremely variable patterns, such as *ACGTACGT*, are never observed, either because they never occur or because they are censored during the alignment procedure. The maximum diversity $\mathcal{N}_s$ that can be observed in an alignment, no matter how long it is, is but a small subset of $\mathcal{A}^s$.

The support $\mathcal{N}_s$ of Q is of course unknown but estimating $N = |\mathcal{N}_s|$ from the distribution of patterns observed in the data set is a classical problem in ecology. We index the patterns of $\mathcal{N}_s$ by (y_j) with j varying from 1 to N and note Y_j the number of occurrences of pattern y_j in the data set. Formally $Y_j = \sum_{i=1}^n 1_{\{\mathbf{X}_i = y_j\}}$ and $(Y_1, \ldots, Y_N)$ is a multinomial random variable $\mathcal{M}(n, \theta^{y_1}, \ldots, \theta^{y_N})$. Note $c = \sum_j 1_{\{Y_j \neq 0\}}$ the number of patterns observed once or more in the data set and $c_i = \sum_j 1_{\{Y_j = i\}}$ the number of patterns observed exactly i times. We have $N = c(1 + c_0/c)$. If the odd $\gamma = c_0/c$ that a pattern goes undetected can be estimated by $\hat{\gamma}$, N can be estimated by $c(1 + \hat{\gamma})$.

Several such estimators come from species abundance problem in ecology. Chao (1984) offers the following estimator: $\hat{\gamma} = c_1^2/cc_2$. The rationale behind this estimator is that the ratio of unobserved patterns to patterns observed exactly 1 time should be similar to the ratio of observed 1 time to observed 2 times. Unfortunately this estimator gives extremely high values of $\hat{\gamma}$. Mao and Lindsay (2007) offer a more involved estimator based on slightly different premises. They assume that the Y_i are independent Poisson variables of parameters λ_i and that the $lambda_i$ arise from a mixture density R on $(0, \infty)$. Under this framework, Y_i is a R-poisson mixture distribution with density f_R and the odd γ that a class goes unobserved is $\gamma = f_R(0)/(1 - f_R(0))$. Conditional on c, the positive Y_i are a c-sample of 0-truncated Poisson mixture with density $f_S(x) = f_R(x)/(1 - f_R(0)) = \int \lambda^x/(x!(e^\lambda - 1))dS(\lambda)$ where, $dS(\lambda) = (1 - e^{-\lambda})dR(\lambda)/\int(1 - e^{-\lambda})dR(\lambda)$ is a simple transformation of R with no mass on 0. Note also that γ is a function of S as $\gamma = f_S(0) = \int(e^\lambda - 1)^{-1}dQ(\lambda)$ so that S (and not R) is sufficient to calculate γ. Mao and Lindsay (2007) strategy builds upon this remark and first estimate S from all distributions with no mass on 0 by $\hat{S}$ using the empirical distribution of the positive Y_i before estimating γ from the $\hat{S}$. Formally, their estimator is defined as follows:

$$\hat{\gamma} = \inf\{\gamma(S) : d(F_S, \hat{F}_n) < \varepsilon_n, S \in \mathcal{F}\}$$

where d is the Kolmogorov distance, ε_n is the $1-\alpha$ quantile of the Kolmogorov distance between uniform $(0, 1)$ and its empirical version over a sample of size n, $\hat{F}_n(x) = \sum_{1 \le i \le x} \frac{n_i}{n}$ and $F_S(x) = \sum_{1 \le i \le x} f_S(i)$. $\mathcal{F}$ is the set of all distribution on $\mathbb{R}$ with no mass on 0, f_S is the S-mixture of Poisson density with no mass on 0 defined $f_S(x) = \int \lambda^x/(x!(e^\lambda - 1))dS(\lambda)$ and $\gamma(S) = \int(e^\lambda - 1)^{-1}dS(\lambda)$. The probability that $|\mathcal{N}_s|$ is greater than this value is at most α. In practice, Mao and Lindsay (2007) suggests discretizing the problem by choosing a grid in $\mathbb{R}_+$ and minimizing over distributions on this grid. The problem can then be solved by linear programming.

Lemma 9 *In a way similar to $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_n$, let $\mathbf{log}\, \boldsymbol{P^m}$ be the vector of size $|\mathcal{N}_s|$ defined by $\mathbf{log}\, \boldsymbol{P^m} = (\log \mathbb{P}(x, m))_{x \in \mathcal{N}_s}$. Then:*

$$\ell^m = \ell_n^m + (\boldsymbol{\theta} - \boldsymbol{\theta}_n)'.\, \mathbf{log}\, \boldsymbol{P^m} \tag{A.3}$$

With ℓ^m defined as in Def. 5, classical properties of the KL-distance ensures that maximizing ℓ^m over models m is equivalent to minimizing the KL-distance between $\mathbb{P}(.; m)$, the pattern distribution induced by model m, and Q, the true one (Kullback and Leibler, 1951).

The true log-likelihood ℓ^m is the sum of two quantities: the computable *empirical* log-likelihood ℓ_n^m and the unknown correction term $(\boldsymbol{\theta} - \boldsymbol{\theta}_n)'.\, \mathbf{log}\, \boldsymbol{P^m}$. To control the difference $\ell^m - \ell_n^m$, the model m is not enough, we also need information on the difference $\boldsymbol{\theta} - \boldsymbol{\theta}_n$.

A.2.3 Distance between Q and Q_n

$\boldsymbol{\theta} - \boldsymbol{\theta}_n$ is a random vector of dimension $|\mathcal{N}_s|$ whose components sum up to 0 and which fluctuates around 0. Our goal here is to bound the probability of this vector being "large", *i.e* away from 0. The component x of $\boldsymbol{\theta} - \boldsymbol{\theta}_n$ is $Y_x^n/n - \theta_x$ where Y_x^n is a binomial $\mathcal{B}(n, \theta_x)$, it is thus centered with variance $(1-\theta_x)\theta_x/n$. As it is fairly easy to obtain concentration inequalities for binomials, we first work component by component before dealing with the complete vector and then concluding on $\ell^m - \ell_n^m$.

Lemma 10 *Let $\boldsymbol{\theta} = (\theta^x)_{x\in\mathcal{N}_s}$ and $|\boldsymbol{\theta}| = (|\theta^x|)_{x\in\mathcal{N}_s}$. Let $\boldsymbol{\varepsilon} = (\varepsilon_x)_x$ be a vector with positive components, then:*

$$\mathbb{P}\left(\bigcup_{x\in\mathcal{N}_s}\left\{|\theta^x - \theta_n^x| > \varepsilon_x\right\}\right) \leq |\mathcal{N}_s| \max_{x\in\mathcal{N}_s} \mathbb{P}(|\theta^x - \theta_n^x| > \varepsilon_x)$$

The total number of observable patterns, $|\mathcal{N}_s|$, plays a crucial role in the formula as a multiplicative factor of the probability. Hence the need for accurate lower bound of this number, such as those provided by Chao (1984) or Mao and Lindsay (2007). Although it is quite clear that $|\mathcal{N}_s|$ increases with the number s of species, the shape of the increase is not straightforward and depends strongly on the relatedness of the new species to those in the sample. In the extreme case where an additional species is completely similar to one of those in the sample, $|\mathcal{N}_{s+1}| = |\mathcal{N}_s|$. However if the added species is extremely distant from each one of those in the sample, $|\mathcal{N}_{s+1}|$ can be up to 4 times greater than $|\mathcal{N}_s|$.

Consider a sequence (X_n) of i.i.d. Bernoulli variables with parameter p. For a given pattern, X_n represent the presence/absence of the pattern at position n. We upper-bound the probability of $\{|\sum_{i=1}^n (X_i - p)| > n\varepsilon\}$ by bounding the probability of the right-end tail $\{> n\varepsilon\}$ and left-end tail $\{< n\varepsilon\}$.

Lemma 11 *Consider (X_n) a sequence of i.i.d. Bernoulli variables with parameter p. For all $\varepsilon \geq 0$, we have:*

$$\log P\left(\sum_{i=1}^n (X_i - p) > n\varepsilon\right) \leq \frac{-n\varepsilon^2}{2p(1-p)}\left[1 - \frac{\varepsilon}{6p(1-p)}\right]$$

and

$$\log P\left(\sum_{i=1}^n (X_i - p) < -n\varepsilon\right) \leq \frac{-n\varepsilon^2}{2p(1-p)}\left[1 - \frac{\varepsilon}{6p(1-p)}\right]$$

The proof of Lemma 9 is postponed to the appendix. Large deviations theory (Dembo and Zeitouni, 1998) tells us that the probability of the unlikely event $\{|\sum_{i=1}^n (X_i - p)| > n\varepsilon\}$ decays exponentially with n. The main purpose of this Lemma is to uncover the exponential speed (right-hand side of the equations).

As the probability of observing a pattern at a given site is usually quite small (with the notable exception of invariant patterns), p is usually much smaller than $1-p$. We are not so much interested in the absolute deviation of $\boldsymbol{\theta}_n$ from $\boldsymbol{\theta}$ rather than in the relative deviation. $\boldsymbol{\varepsilon}$ is thus chosen as a fraction of $\boldsymbol{\theta}$ so that the ratio ε_x/θ^x is small.

Proposition 12 *For ε_x/θ^x smaller than $a < 1$ and $\max_{x\in\mathcal{N}_s} \theta^x \leq 5/6$, we have:*

$$\log \mathbb{P}\left(\bigcup_{x\in\mathcal{N}_s} \left\{|\theta^x - \theta_n^x| > \varepsilon_x\right\}\right) \leq \log|\mathcal{N}_s| + \log 2 + \max_{x\in\mathcal{N}_s} \frac{-n(1-a)\varepsilon_x^2}{2\theta^x(1-\theta^x)} \tag{A.4}$$

Proof. The two end tails of $\theta^x - \theta_n^x$ are bounded as in Lemma 9. With the conditions on the θ^x and the ε_x, remark that $\frac{\varepsilon_x}{6\theta^x(1-\theta^x)} \leq a < 1$ so that:

$$\log \mathbb{P}(|\theta^x - \theta_n^x| > \varepsilon_x) < \log 2 - \frac{n(1-a)\varepsilon_x^2}{2\theta^x(1-\theta^x)}$$

Combining this with the result of 8 gives the result. ■

The θ^x giving the smallest decreasing rate for the exponential bound are those close to 1/2. We can explicitly compute the worst rate associated with this value but even the most frequent patterns, the invariant ones, have a frequency nowhere near 1/2. We expect most, if not all, of the θ^x to be much smaller than 1/2 and thus the decreasing rate to be significantly higher.

A.3 Phylogenetic reconstruction for finite size samples

In the following, we take advantage of Prop. 11 on $\boldsymbol{\theta} - \boldsymbol{\theta}_n$ to bound the difference $\ell^m - \ell_n^m$. After doing so, we focus on inversions probabilities, *i.e.* incongruities between the empirical likelihood ranking and the true one.

A.3.1 Distance between the empirical and true mean log-likelihoods

In this part, the goal is to evaluate the confidence given to the log-likelihood of a tree. The smaller this confidence, the more caution is required when dealing with that log-likelihood, for example when comparing it for two trees. To do so, we connect $\ell^m - \ell_n^m$ to $\boldsymbol{\theta} - \boldsymbol{\theta}_n$ using Eq. (A.3).

Corollary 13 *Let $\varepsilon \geq 0$ and note $\tilde{\varepsilon} = \dfrac{\varepsilon}{|\mathcal{N}_s| \times \|\log \boldsymbol{P}^m\|_\infty}$. Let a be the θ^x closest to 1/2. Then:*

$$\log \mathbb{P}\left(|\ell^m - \ell_n^m| \geq \varepsilon\right) \leq \log|\mathcal{N}_s| + \log 2 + \frac{-n\tilde{\varepsilon}^2/2}{a(1-a) + \tilde{\varepsilon}/3} \tag{A.5}$$

Proof. Since

$$|\ell^m - \ell_n^m| = |(\boldsymbol{\theta}_n - \boldsymbol{\theta})'.\log \boldsymbol{P}^m| \leq |\mathcal{N}_s| \times \|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|_\infty \times \|\log \boldsymbol{P}^m\|_\infty,$$

we have

$$\begin{aligned}\mathbb{P}\left(|\ell^m - \ell_n^m| \geq \varepsilon\right) &\leq \mathbb{P}\left(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}\|_\infty \geq \tilde{\varepsilon}\right) \\ &= \mathbb{P}\left(\bigcup_{x\in\mathcal{N}_s} \{|{\theta^x}_n - \theta^x| \geq \tilde{\varepsilon}\}\right) \\ &\leq |\mathcal{N}_s| \max_{x\in\mathcal{N}_s} \mathbb{P}\left(|{\theta^x}_n - \theta^x| \geq \tilde{\varepsilon}\right) \\ &\leq 2|\mathcal{N}_s| \max_{x\in\mathcal{N}_s} \exp\left\{\frac{-n\tilde{\varepsilon}^2/2}{\theta^x(1-\theta^x) + \tilde{\varepsilon}/3}\right\}\end{aligned}$$

where the last inequality is a direct application of Bernstein's inequality. Since $\frac{-n\tilde{\varepsilon}^2/2}{\theta^x(1-\theta^x)+\tilde{\varepsilon}/3}$ is symmetric around and reaches its maximum at $\theta^x = 1/2$, it is maximized by $\theta^x = a$, the closest to $1/2$.

■

The decaying rate obtained is Eq. (A.5) is driven by a. Since most patterns have small occurrence probability ($\ll 1/2$), a turns out to be the probability of the most frequent pattern. $|\mathcal{N}_s|$ is the number of possible patterns and appears twice in the formula, once as a multiplicative factor $\log|\mathcal{N}_s|$ and once as a rescaling of the deviation ε. Since $|\mathcal{N}|_s$ is unknown, it is replaced with an estimator. The more accurate our estimator, the finer our inequality is. For example, dividing the bound by 2 gives a multiplicative factor twice smaller and a decreasing rate four times faster. Cor. 12 controls the absolute deviation of ℓ_n^m from ℓ^m. A similar looking inequality can be obtained for the relative deviation, allowing for statement like: "with probability greater than 0.95, ℓ_n^m is between $(1-\alpha)\ell^m$ and $(1+\alpha)\ell^m$".

Corollary 14 *Consider $\alpha \in (0,1)$ and assume than $\max_{x\in\mathcal{N}_s} \theta^x \leq 5/6$. Then:*

$$\log \mathbb{P}\left(\left|\frac{\ell^m - \ell_n^m}{\ell^m}\right| \geq \alpha\right) \leq \log|\mathcal{N}_s| + \log 2 + \max_{x\in\mathcal{N}_s} \frac{-n(1-\alpha)\alpha^2\theta^x}{2(1-\theta^x)} \tag{A.6}$$

Proof. If each component of $\boldsymbol{\theta} - \boldsymbol{\theta}_n$ is within a factor α of $\boldsymbol{\theta}$, then $\ell^m - \ell_n^m$ is also within a factor α of ℓ^m so that:

$$\mathbb{P}\left(|\ell^m - \ell_n^m| \geq \alpha|\ell^m|\right) \leq \mathbb{P}\left(\bigcup_{x\in\mathcal{N}_s} \left\{|\theta^x - \theta_n^x| \geq \alpha\theta^x\right\}\right)$$

Replacing ε_x by $\alpha\theta^x$ in Eq. (A.4) then provides:

$$\mathbb{P}\left(|\boldsymbol{\theta} - \boldsymbol{\theta}_n| \geq \alpha\boldsymbol{\theta}\right) \leq \log|\mathcal{N}_s| + \log 2 + \max_{x\in\mathcal{N}_s} \frac{-n(1-\alpha)\alpha^2\theta^x}{2(1-\theta^x)}$$

■

In Cor. 13, the exponential decay is limited by those patterns whose probability is low. It can be understood easily: estimating ℓ^m to a relative precision α requires

estimation of each θ^x to a precision $\alpha\theta^x$, implying that the accuracy needed for low frequency patterns is really high. A strategy consisting of treating separately those sites with very low frequencies (with respect to the sample size) and the other ones might give higher decay but is not explored here.

As hinted by Eq. (A.3), two patterns are different for our purpose only so far as they account for different likelihood values under model m. Depending on the model and the topology, we can replace $|\mathcal{N}_s|$ by an even smaller value. Simple models generate only a few patterns while more complicated models generate more patterns. For example, under a *K2P* model and a quartet tree *AB|CD* (*i.e.* A and B are separated from C and D, see Fig. A.1), out of the $4^4 = 256$ different patterns, there are only 30 different values for the probability of occurrence of a pattern.

A.3.2 Support given to a tree

A further step is the ranking on models induced by their mean log-likelihood. Since the true log-likelihoods of models are not achievable, rankings are based upon their empirical log-likelihoods. Of course, inversion events can happen: when comparing two models m and m', the empirical log-likelihoods could by chance give a different ranking than the true one. Since the maximum *empirical* log-likelihood model is retrieved, this is unwanted. We offer here to bound the probability of such an event.

Proposition 15 *Assume that model m is better than model m' in the sense that $\ell^m > \ell^{m'}$. Then, the probability that m' is better than m for the sample: $\mathbb{P}\left(\ell_n^m - \ell_n^{m'} < 0\right)$ is such that:*

$$\log \mathbb{P}\left(\ell_n^m - \ell_n^{m'} < 0\right) \leq \log |\mathcal{N}_s| + \max_{x \in \mathcal{N}_s} \frac{-n\theta^x (1-\varepsilon)\,\varepsilon^2}{2(1-\theta^x)} \tag{A.7}$$

$$\text{where } \varepsilon = \frac{\ell^m - \ell^{m'}}{\|\log \boldsymbol{P^m} - \log \boldsymbol{P^{m'}}\|_\infty} \leq 1$$

Proof. Since $\ell_n^m - \ell_n^{m'} = \frac{1}{n}\sum_{i=1}^n \log \mathbb{P}(X_i; m) - \log \mathbb{P}(X_i; m')$, we can use Lemma 9 to bound $\ell_n^m - \ell_n^{m'} - (\ell^m - \ell^{m'})$ in the same way than $\ell_n^m - \ell^m$. We just need to replace $\|\log \boldsymbol{P^m}\|_\infty$ by $\|\log \boldsymbol{P^m} - \log \boldsymbol{P^{m'}}\|_\infty$.

$$\begin{aligned}
\Delta &= \mathbb{P}\Big(\ell_n^m - \ell_n^{m'} < 0\Big) \\
&= \mathbb{P}\Big(\ell_n^m - \ell_n^{m'} - (\ell^m - \ell^{m'})) < -(\ell^m - \ell^{m'})\Big) \\
&\leq \mathbb{P}\left(\bigcup_{x \in \mathcal{N}_s} \left\{(\theta_n^x - \theta^x) \log \frac{\mathbb{P}(x;m)}{\mathbb{P}(x;m')} < -\theta^x(\ell^m - \ell^{m'})\right\}\right) \\
&\leq |\mathcal{N}_s| \max_{x \in \mathcal{N}_s} \mathbb{P}\left((\theta_n^x - \theta^x) \log \frac{\mathbb{P}(x;m)}{\mathbb{P}(x;m')} < -\theta^x(\ell^m - \ell^{m'})\right)
\end{aligned}$$

Keep in mind that $\ell^m - \ell^{m'} > 0$. If $\log \dfrac{\mathbb{P}(x;m)}{\mathbb{P}(x;m')} > 0$,

$$\begin{aligned}
\mathbb{P}\left((\theta_n^x - \theta^x)\log\frac{\mathbb{P}(x;m)}{\mathbb{P}(x;m')} < -\theta^x(\ell^m - \ell^{m'})\right) &= \mathbb{P}\left(\theta_n^x - \theta^x < -\theta^x\frac{\ell^m - \ell^{m'}}{\log\mathbb{P}(x;m) - \log\mathbb{P}(x;m')} < 0\right) \\
&\leq \mathbb{P}\left(\theta_n^x - \theta^x < -\theta^x\frac{\ell^m - \ell^{m'}}{\|\log \boldsymbol{P^m} - \log \boldsymbol{P^{m'}}\|_\infty} < 0\right) \\
&\leq \exp\left\{\frac{-n(1-\varepsilon)\varepsilon^2\theta^x}{2(1-\theta^x)}\right\}
\end{aligned}$$

where the last inequality stems from Lemma 9. If $\log \dfrac{\mathbb{P}(x;m)}{\mathbb{P}(x;m')} < 0$, the same argument yields

$$\begin{aligned}
\mathbb{P}\left((\theta_n^x - \theta^x)\log\frac{\mathbb{P}(x;m)}{\mathbb{P}(x;m')} < -\theta^x(\ell^m - \ell^{m'})\right) &= \mathbb{P}\left(\theta_n^x - \theta^x > \theta^x\frac{\ell^m - \ell^{m'}}{\log\mathbb{P}(x;m') - \log\mathbb{P}(x;m)} > 0\right) \\
&\leq \mathbb{P}\left(\theta_n^x - \theta^x > \theta^x\frac{\ell^m - \ell^{m'}}{\|\log \boldsymbol{P^m} - \log \boldsymbol{P^{m'}}\|_\infty} > 0\right) \\
&\leq \exp\left\{\frac{-n(1-\varepsilon)\varepsilon^2\theta^x}{2(1-\theta^x)}\right\}
\end{aligned}$$

Finally, if $\mathbb{P}(x;m) = \mathbb{P}(x;m')$,

$$\mathbb{P}\left((\theta_n^x - \theta^x)\log\frac{\mathbb{P}(x;m)}{\mathbb{P}(x;m')} < -\theta^x(\ell^m - \ell^{m'})\right) = 0 \leq \exp\left\{\frac{-n(1-\varepsilon)\varepsilon^2\theta^x}{2(1-\theta^x)}\right\}$$

Wrapping everything together gives the result. ■

Remark: This result is expected: the farther ℓ^m and $\ell^{m'}$ are, the less likely inversion events are. As for Cor. 12, $|\mathcal{N}_s|$ can be reduced: indeed patterns x equally supporting models m and m', *i.e* satisfying $\log\mathbb{P}(x;m) = \log\mathbb{P}(x;m')$ can be discarded as they do not contribute to $\ell_n^m - \ell_n^{m'}$. For simple substitution models, there is a fair number of such patterns.

The bound derived in Prop. 14 relies on two a priori unknown quantities: the number of patterns $|\mathcal{N}_s|$ and the difficulty of the problem, given by the factor $\Delta = \dfrac{(\ell^m - \ell^{m'})^2}{\|\log \boldsymbol{P^m} - \log \boldsymbol{P^{m'}}\|^2} \times \min_{x\in\mathcal{N}_s} \dfrac{\theta^x}{2(1-\theta^x)}$. $|\mathcal{N}_s|$ can be estimated from the number of patterns observed in the sample using the results from Mao and Lindsay (2007). Δ is the product of two ratios. The first one takes value in $[0;1]$ and reflects the relatedness of two models: 1 means that the two models are as different as can be whereas 0 means that they have the same likelihood score, and thus are not distinguishable by a likelihood method. The second ratio $\min_{x\in\mathcal{N}_s} \dfrac{\theta^x}{2(1-\theta^x)}$ takes value in $(0;1/2)$ and is driven by the smallest probability pattern: close to 0 when the least likely pattern has a very small occurrence probability are equally distributed and close to $1/2$ when one pattern dominates the distribution. To sum up, Δ can take any value between 0 and $1/2$ and is bound to be somewhere between these two extremes.

Since $\mathbb{P}(\ell_n^m - \ell_n^{m'} < 0) \leq |\mathcal{N}_s| \exp(-n\Delta)$, we compute for two given confidence levels, 0.95 and 0.66, any thus the smallest assessable Δ as a function of n and $|\mathcal{N}_s|$, namely:

$$\Delta(n, |\mathcal{N}_s|) = \frac{1}{n} \log \frac{|\mathcal{N}_s|}{\alpha} \tag{A.8}$$

where the confidence level is $1 - \alpha$. Not surprisingly, Δ decreases with n (better accuracy) and increases with $|\mathcal{N}_s|$ (lower accuracy).

Using Cor. 13, we can in the same way compute for a given confidence level $1 - \alpha$, number of sites n and precision β, the value $\xi(\alpha, n, \beta)$ such that all models which smallest probability value is above this value can be scored within a range $(1-\beta, 1+\beta)$ of their true value.

$$\xi(\alpha, n, \beta) = \frac{1}{n\beta^2} \log \frac{|\mathcal{N}_s|}{\alpha} \tag{A.9}$$

For example, for 20 patterns the minimum Δ ranges from $2.99e^{-2}$ (resp. $2.05e^{-2}$) for 200 sites to $2.99e^{-3}$ (resp. $2.05e^{-3}$) for 2500 sites for the 95% confidence level (resp. 66% level). For 100 patterns, the minimum Δ ranges from $3.8e^{-2}$ (resp. $2.85e^{-2}$) for 200 sites to $3.8e^{-3}$ (resp. $3.85e^{-3}$) for 2500 sites for the 95% confidence level (resp. 66% level). However, for 100 patterns, the smallest probability θ^x may be much lower than for 20 patterns so that the models need to be more separated (higher $|\ell^m - \ell^{m'}|$ for Δ to remain constant. The same values for ξ gives: for 20 patterns and with precision 5%, the minimum ξ ranges from 0.06 (resp. 0.041) for 200 sites to $3.83e^{-4}$ (resp. $2.61e^{-4}$) for 2500 sites for the 95% confidence level (resp. 66% level). For 100 patterns, the corresponding values are 0.076 (resp. 0.056) and $4.86e^{-4}$ (resp $3.6e^{-4}$).

Overall the proposed method is suited for small numbers of patterns ($\leq$ 150) induced by either simple evolution models or quite closely related species.

A.4 Comparison with widely used methods and illustration

Several methods already exist in the literature to test the stability of a tree. The most popular is bootstrap, widely used to test two candidate topologies. Cor. (12) and Prop. (14) basically have the same goal: control the fluctuations of a random estimator around its asymptotic value, but we are interested in the mean likelihood of a nucleotide under model m rather than in the tree T embedded in model m. Building confidence intervals on trees is out of the scope of this article (Shimodaira, 2002; Shimodaira and Hasegawa, 1999).

A.4.1 Bootstrap

Bootstrap procedures in phylogeny are based on sampling with replacement in the data (Felsenstein, 1985; Goldman et al., 2000; Holmes, 2003). For example, to test the

stability of a clade present in tree $T^{(0)}$ inferred from the data, draw B bootstrap samples $(\mathcal{X}^{(i)})_{i=1..B}$ and infer the Maximum-Likelihood tree $T^{(i)}$ for each bootstrap sample. Note b the number of $T^{(i)}$ in which the clade of interest is present. The bootstrap support of $T^{(0)}$ is then $(b+1)/(B+1)$. This support is compared to an arbitrary threshold, usually 0.66 or 0.95 (see Felsenstein (2004) chap.20), and if higher, the clade is declared present with a 0.66 or 0.95 support.

On top of the already known and widely discussed problems of bootstrap (Efron et al., 1996; Felsenstein, 2004; Hillis and Bull, 1993; Holmes, 2003; Newton, 1996) it is obvious that setting the threshold *ex ante* has some disadvantages: the major one is to leave out both n and s. On the one hand, for small n, the stochastic effects can be large so that even if the clade is there, it can be absent from bootstrap trees more often than 5% of the times. On the other hand, when n is large enough, the inferred tree has a high probability of having the same topology as the true one, and hence the same clades. In this case, a clade present in the true tree will appear in more than 95% of the bootstrap trees. It is then interesting to set a threshold higher than 0.95 to build a more conservative test. Anyhow, the threshold should include the numbers s of species, the number n of nucleotides and the complexity of the substitution model.

Bootstrap procedures can also be used for testing phylogenies, using the K-H test (Kishino and Hasegawa, 1989). The most popular forms of K-H test relies either on the RELL (resampling estimated log-likelihood) approximation, or a normal approximation of $\ell_n^m - \ell_n^{m'}$. Under the RELL approximation, log-likelihoods of sites (estimated under the best model for the original data set) are resampled instead of sites themselves. Under the normal approximation, variance of the normal approximation is computed from the bootstrap samples and the significance of $\ell_n^m - \ell_n^{m'}$ is evaluated with regards to the normal distribution with estimated variance rather than to the empirical distribution derived from the bootstrap samples. Normal approximation is of course tighter than any concentration inequality can hope to be. However, the normal approximation assumes the evolution model is well specified. Model misspecification is often of great concern as even the most sophisticated evolution model are unable to grasp all the subtleties of molecular evolution. And even if the correct model is time reversible Markov process along the tree, choosing it among the many candidates is not an easy task (Posada and Buckley, 2004). Moreover, unlike concentration bounds which assume nothing on the evolution models of the trees being compared, the correctly specified model assumption prevents one to compare two trees with different evolution models. Last but not least, the K-H test requires the empirical pattern distribution to be a good approximation of the real one, whereas our goal is precisely to study the uncertainty arising from the two being different. To sum up, K-H test provides the user with tight bounds as a counterpart of stringent assumptions.

A caveat of bootstrap common both to bootstrap values and the K-H test, at least under its normal form, is the lack of dependence on n and s. It is quite sensible that the number n of nucleotide required to achieve a given level of confidence on a phylogenetic model over s species depends on s. Namely it grows, possibly quite rapidly, with s. While bootstrap procedures are powerless to calculate this n, we can retrieve it using our analytical techniques. Analytical bounds are also very comfortable as they let us study both convergence speeds and the importance of initial hypothesis on the stability of the mean log-likelihood.

Finally and albeit this concern vanishes as computing power increases, bootstrap has some limitations as it relies heavily on simulations to compute support probabilities. For large values of n and s, the computational burden can be prohibitive. The proposed method upper bounds such probabilities instead of approximating them, but in an analytical way: the computational burden is not a problem anymore.

A.4.2 Illustration of the method on an example

Presentation We introduce here an example, consisting of binary characters and 4 species: A, B, C and D. Binary characters can be thought as purine/pyrimidine. The model is quite trivial as $s = 4$ is quite small and almost any method would be able to retrieve the correct topology, although with different branch lengths and no mixture on the evolution model. Our goal here is not to outperform existing estimation method but rather to show how often and when they fail when the true evolution model is not accounted for.

We assume that the true topology is AB|CD (*i.e.* A and B are separated from C and D, see Fig. A.1) and that the true evolution model is an usual symmetric model but complicated by correlated evolution between A and B for some of the sites: for those sites if a change on the terminal branch leading to A, the same happens on the terminal branch leading to B contrasting with the usual situation where evolution on those branches is usually independent (conditional on the parent node). Formerly, a fraction p of the sites evolve on the left-hand side tree of Fig. A.1 and a fraction $1-p$ on the right-hand side of Fig. A.1. For the sake of simplicity, all terminal branch lengths are equal, we note e the probability of change on a terminal branch of length t_1, f the same probability on the central one of size t_0 and $g = (e + f - 2ef)$ the probability of change on a branch of length $t_0 + t_1$. When the terminal branch lengths are not equal, we need to consider four probabilities $e_1, \ldots, e_4$, one for each terminal branch, instead of just e. The distributions are still tractable but a bit more complicated than the simple case we consider. e, f and g can take any value in $[0, 1/2]$ as the branch lengths vary from 0 to ∞.

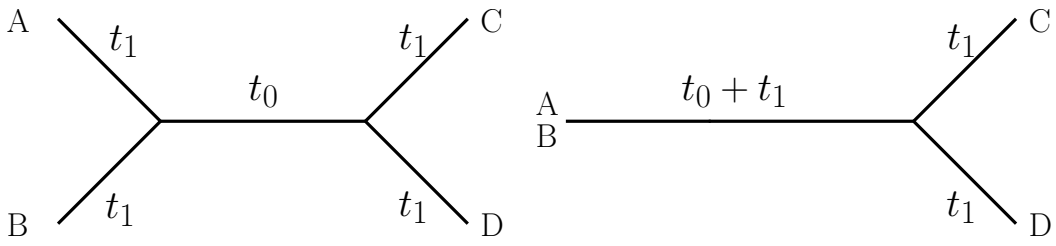

Figure A.1: Left: topology on which evolve the fraction p of sites with no correlated evolution. Right: topology on which evolve the fraction $1 - p$ of sites with correlated evolution.

Since the model is symmetric, there are 8 different patterns, $xxxx$, $yxxx$, $xyxx$, $xxyx$, $xxxy$, $xxyy$, $xyxy$ and $xyyx$ whose probabilities are given in Tab. A.1

As in most four species studies, the interest in determining which of the three topologies T_1, T_2 and T_3 of Fig. A.2 is the best. To do this, we consider the trees with equal terminal branch lengths associated to these three topologies (see Fig. A.2).

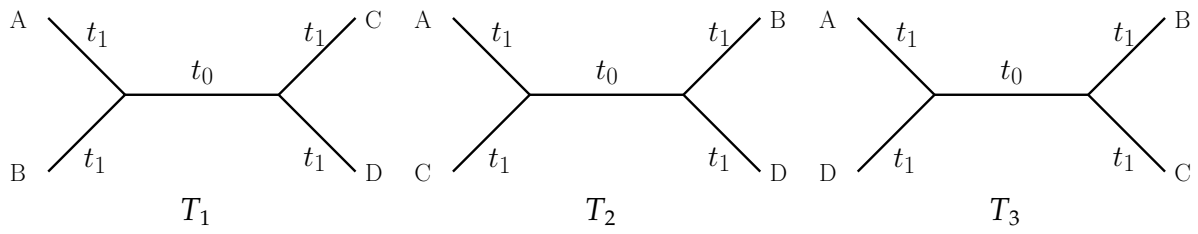

Figure A.2: The three candidates trees

Under the markovian evolution model described in the first paragraph, tree T_i induce a phylogenetic model m_i and a pattern distribution P^{T_i} (we adopt P^{T_i} instead of P^{m_i} to emphasize the interest in the tree). The distributions P^{T_1}, P^{T_2} and P^{T_3} are very similar to each other and to the distribution under no correlated evolution. The only difference between them lies in the probabilities of patterns $xxyy$, $xyxy$ and $xyyx$ (see Tab. A.2) which basically give some information on the position of the central branch. Note that because of the mixture component of the real evolution model, none of the P^{T_i} can reproduce the real distribution Q, such that the likelihoods computed under a P^{T_i} when they should be computed under Q are slightly off.

With these distributions, it is easy to calculate the likelihood scores of each candidate tree. For $e = 0.15$, $f = 0.20$ and $p = 0.8$ (branch length typical of the placental mammals phylogeny and correlated evolution in 20% of the sites), the expressions of Tab. A.1 and A.2 give:

	$xxxx$	$yxxx$	$xyxx$	$xxyx$	$xxxy$	$xxyy$	$xyxy$	$xyyx$
Q	44.4	7.6	7.6	10.1	10.1	15.0	2.6	2.6
P^{T_1}	42.5	9.5	9.5	9.5	9.5	13.1	3.3	3.3
P^{T_2}	42.5	9.5	9.5	9.5	9.5	3.3	13.1	3.3

where the probabilities are expressed as percentages. Using theses values, we compute the likelihood scores of the candidate trees according to Eq. (A.2):

$$\ell^{T_1} = -1.70 \quad \text{and} \quad \ell^{T_2} = \ell^{T_3} = -1.87$$

Table A.1: Pattern distribution with no correlated evolution (P_1, left tree of Fig. A.1), only correlated evolution (P_2, right tree of Fig. A.1) and for the mixture of both (Q).

	Probability of the pattern		
Pattern	P_1(Pattern)	P_2(Pattern)	Q(Pattern)
$xxxx$	$e^4 + (1-e)^4 - f(1-2e)^2$	$(1-e)^2(1-g) + e^2 g$	$pP_1(xxxx) + (1-p)P_2(xxxx)$
$yxxx$	$e(1-e)(1-2e(1-e))$	0	$pP_1(yxxx)$
$xyxx$	$e(1-e)(1-2e(1-e))$	0	$pP_1(yxxx)$
$xxyx$	$e(1-e)(1-2e(1-e))$	$e(1-e)$	$(1-e)(1-2pe(1-e))$
$xxxy$	$e(1-e)(1-2e(1-e))$	$e(1-e)$	$(1-e)(1-2pe(1-e))$
$xxyy$	$f(1-2e)^2 + 2e^2(1-e)^2$	$g(1-e)^2 + (1-g)e^2$	$pP_1(xxyy) + (1-p)P_2(xxyy)$
$xyxy$	$2e^2(1-e)^2$	0	$pP_1(xyxy)$
$xyyx$	$2e^2(1-e)^2$	0	$pP_1(xyxy)$

Table A.2: Pattern distribution under the three candidates trees: T^1, T^2 and T^3. α stands for $P_1(xxyy)$ and β for $P_1(xyxy)$.

	Probability of the pattern		
Pattern	P^{T_1}(Pattern)	P^{T_2}(Pattern)	P^{T_3}(Pattern)
$xxyy$	α	β	β
$xyxy$	β	α	β
$xyyx$	β	β	α

We now consider bounds on the $\ell_n^{T_1} - \ell^{T_1}$. Before doing this, we need to consider the effective number of patterns $|\mathcal{N}_s|$. Although 8 patterns are considered, they account for only 4 log-likelihood values under P^{T_1}: $\log(0.425)$, $\log(0.095)$, $\log(0.131)$ and $\log(0.033)$. We can thus merge the corresponding patterns under super patterns: $xxxx$ and $xxyy$ remain unchanged, $yxxx$, $xyxx$, $xxyx$ and $xxxy$ are merged into a super-pattern, noted $yxxx$, while $xyxy$ and $xyyx$ are also merged in another super-pattern, noted $xyxy$. Q is modified accordingly:

	$xxxx$	$yxxx$	$xxyy$	$xyxy$
Q	44.4	$2 \times 7.6 + 2 \times 10.1 = 35.5$	15.0	$2 \times 2.6 = 5.2$

The support of Q is reduced to 4 patterns, thus $|\mathcal{N}_s| = 4$. Using Cor. 12 and 13, we need both the smallest probability and the closest to 1/2, respectively 0.052 and 0.444. The pattern determining $\|\log \boldsymbol{P}^{\boldsymbol{T_1}}\|_\infty$ is the pattern (not the super-pattern) of smaller probability under P^{T_1}: $xyxy$ which gives $\|\log \boldsymbol{P}^{\boldsymbol{T_1}}\|_\infty = \log(0.026) = 3.42$. With all these quantities, we can compute the exponential decay rate for $\varepsilon = 0.1$ (absolute precision of 0.1 for Cor. 12 and relative precision of 10% for Cor. 13), respectively:

$$\begin{aligned}
\tilde{\varepsilon} = \frac{\varepsilon}{|\mathcal{N}_s| \|\log \boldsymbol{P}^{\boldsymbol{T_1}}\|_\infty} &= \frac{0.1}{4 \times 3.42} = 7.31e^{-3} \\
-\frac{\tilde{\varepsilon}^2}{2} \min_{x \in \mathcal{N}_s} \frac{1}{\theta^x(1-\theta^x) + \tilde{\varepsilon}/3} &= \frac{-(7.31e^{-3})^2/2}{0.444(1-0.444) + 7.31e^{-3}/3} = -1.07e^{-4} \\
-\varepsilon^2 \min_{x \in \mathcal{N}_s} \frac{\theta^x}{1-\theta^x} &= -0.1^2 \times \frac{0.052}{1-0.052} = -5.49e^{-4}
\end{aligned}$$

We now use Prop. 14 to bound the probability of ranking uncorrectly ℓ^{T_1} and ℓ^{T_2}. What matters now is $\log \boldsymbol{P}^{\boldsymbol{T_1}} - \log \boldsymbol{P}^{\boldsymbol{T_2}}$. For all patterns but $xxyy$ and $xyxy$, the components of $\log \boldsymbol{P}^{\boldsymbol{T_1}} - \log \boldsymbol{P}^{\boldsymbol{T_2}}$ are 0. Except for these two patterns, we can thus discard all patterns from the analysis as they do not participate in the log-likelihood difference. In this case, $|\mathcal{N}_s| = 2$. From the previous computations $\ell^{T_1} - \ell^{T_2} = 0.17$ and $\|\log \boldsymbol{P}^{\boldsymbol{T_1}} - \log \boldsymbol{P}^{\boldsymbol{T_2}}\| = \log P^{T_1}(xxyy) - \log P^{T_2}(xxyy) = \log(0.131/0.033) = 1.39$ and the smallest pattern probability from $Q(xxyy)$ and $Q(xyxy)$ is 0.052. We thus have $\Delta = 0.17^2/1.39^2 \times 0.052(1-0.052)^{-1} = 8.2e^{-4}$ requiring 4500 sites to correctly identify the best tree.

Proof of the lemmas

We prove here Lemmas 9

Proof of lemma 9

Proof. We prove only the first inequality. The second one is deduced from the first one by changing X_i to $1 - X_i$ and p to $1 - p$. With the notation of Lemma 9, we have (see equation (2.8) of Massart (2006) for a demonstration) for all $\varepsilon \in [0, 1 - p)$:

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - p) > n\varepsilon\right) \leq \exp\left(-nh_p(\varepsilon)\right)$$

where $h_p(\varepsilon) = (1-p-\varepsilon)\log\frac{1-p-\varepsilon}{1-p} + (p+\varepsilon)\log\frac{p+\varepsilon}{p}$. We use two well-known inequalities. For all $x \geq 0$

$$\log(1+x) \geq x - \frac{x^2}{2}$$

And for all $x \in [0, 1)$:

$$\log(1-x) = -\sum_{i=1}^{\infty}\frac{x^i}{i} \geq -x - \frac{x^2}{2} - \frac{x^3}{3(1-x)}$$

So that:

$$\begin{aligned}
h_p(\varepsilon) &= (1-p-\varepsilon)\log\left(1 - \frac{\varepsilon}{1-p}\right) + (p+\varepsilon)\log\left(1 + \frac{\varepsilon}{p}\right) \\
&\geq \frac{\varepsilon^2}{2(1-p)} + \frac{\varepsilon^2}{2p} + \frac{\varepsilon^3}{2(1-p)^2} - \frac{\varepsilon^3}{2p^2} - \frac{\varepsilon^3}{3(1-p)^2} \\
&= \frac{\varepsilon^2}{2p(1-p)}\left(1 + \varepsilon\frac{3(1-p)^2 - p^2}{6p(1-p)}\right) \\
&\geq \frac{\varepsilon^2}{2p(1-p)}\left(1 - \frac{\varepsilon}{6p(1-p)}\right)
\end{aligned}$$

For $\varepsilon \geq 1 - p$, the left hand side is $-\infty$ and the inequality still holds. ■

Appendix B

Assessing the Distribution Consistency of Sequential Data

This section is a modified version of the article *Assessing the Distribution Consistency of Sequential Data* by M. Mariadassou and A. Bar-Hen.

Abstract Given n observations, we study the consistency of a batch of k new observations, in terms of their distribution function. We propose a non-parametric, non-likelihood test based on Edgeworth expansion of the distribution function. The keypoint is to approximate the distribution of the $n + k$ observations by the distribution of $n - k$ among the n observations. Edgeworth expansion gives the correcting term and the rate of convergence. We also study the discrete distribution case, for which Cramèr's condition of smoothness is not satisfied. The rate of convergence for the various cases are compared.

Keywords: Edgeworth expansion, Consistency, Resampling

B.1 Introduction

Let $\mathbf{X} = X_1, \ldots, X_n$ be independent observations from a repeated experiment, and with common distribution function F. Let F_n be the empirical distribution and $S(X_1, \ldots, X_n) = S(F_n)$ be a statistic of the observations. The precision of $S(F_n)$ is a strictly decreasing function of n and the sample size is thus a crucial issue.

It is often possible to increase the sample size by acquiring additional observations $\mathbf{X}' = X_{n+1}, \ldots, X_{n+k}$. This is done at additional cost and time, for example by increasing the cohorts in clinical trials or sequencing additional genes in molecular biology. In a parametric framework where F belongs to some family $(F_\theta)_{\theta\in\Theta}$, $S(X_1, \ldots, X_n)$ would typically be an estimator of θ satisfying $S(F_\theta) = \theta$ and the precision, often of order $n^{-1/2}$, should decrease by using $\mathbf{X}'$. However the truth is often more complex. The use of additional observations raises at least two issues, which are addressed in this paper. The first one is the relevance of additional observations to the inference problem. If the additional observations $\mathbf{X}'$ do not share the distribution function F with $\mathbf{X}$, it is certainly unwise to expect better precision when using them in the inference. We therefore need to assess whether $\mathbf{X}'$ is distributed consistently with F. Focusing on the average modification induced by extending the sample to $\mathbf{X}'$, we provide in Section B.3 an approximation to the law of this modification, under the consistency hypothesis. This approximation is then fed in Section B.4 to a test procedure and used to control the type I error. The second issue is the relevance of acquiring the data. If the common distribution F' of observations in $\mathbf{X}'$ is close to F, one additional observation only is likely not to be enough to detect the difference between F and F'. Indeed k needs to be larger than some function of n for the test to be powerful. In test language, for given F' and F, it is similar to finding the size sample needed to achieve a power exceeding some threshold. This issue can be solved using results of Section B.3 and is addressed in Section B.4.

These two issues arise in a slightly different form in sequential tests of hypotheses and sequential change point detection. When collecting new observations is lengthy and costly, waiting for completion of a sample of size n before performing the analyses is not a option. In such an instance, it is desirable to use any new observation as soon as it becomes available. Wald's Sequential Probability Ratio Test (SRPT), introduced by his seminal paper (Wald, 1945) and tightly connected to the classical Neyman-Pearson test for fixed sample size, does just this. Sequential tests stop sampling as soon as a positive result is detected and can thus be superior to classical tests by providing

results faster than classical tests, as the success story of the Beta-Blocker Heart Attack Trial (BHAT) proved in 1981 when it ended 8 months earlier than scheduled with positive results (Study, 1981).

But, although modifications exists to account for account for composite hypothesis (Brodsky and Darkovsky, 2005), sequential tests usually test $H_0 : F = F_0$ against $H_1 : F = F_1$, *i.e.* observations are either all distributed according to F_0 or all distributed according to F_1, which is different of our main concern, since new data can have a different distribution function than the previous ones. Sequential change point detection is closer in essence to our needs, although it does not perfectly fits our need either.

Sequential change point detection is heavily used in statistical quality control. It is used to answer three questions: has a production process ran out of control, when did it ran out of control and what is the magnitude of the change ? Assume that the observations are distributed according to F_0 under the state of control and according to F_1 under the other state.

Noting T the point in time at which the jump is detected and ν the point at which it occurs, most of the change point detection literature is interested in minimizing $E[(T - \nu)^+]$, the average number of additional observations needed to detect the change. This is very close to our concern: new observations not being consistent with the previous ones is equivalent to a process running out of control at time n. The CUSUM (cumulative sum) charts use the current observation to detect significant departures of the process from the state of control Page (1954). Lai (1995) showed that a moving average scheme consisting of only a finite size observation window around the current observation is asymptotically as efficient as the CUSUM if the window size grows suitably fast to infinity. Brodsky and Darkovsky (2000) generalize this result to a larger class of schemes. But all these methods are likelihood-based and assume F_0 and F_1 are simple enough for log-likelihood ratio to be easily computed. Benveniste et al. (1987) use weak convergence theory to extend CUSUM to non-likelihood-based procedures. Their asymptotic local approach use convergence of the rescaled sums of detection statistics to a gaussian process. Lai and Shan (1999) use another approach based on moderate deviations to extend a Generalized Likelihood Ratio (GLR) to non-likelihood-based detection statistics. We present in this paper an original non-likelihood based method to check the consistency of a new batch of observations with previous ones. Our method requires very little assumption about F_0 and F_1 and builds upon a simple and intuitive idea: under the hypothesis of consistency, the precision gain obtained when adding k observations to the sample can roughly be estimated by the precision loss induced by removing k observations from the sample.

Our work is motivated by the study of DNA sequences. Organisms genomes are sequenced gene by gene: when new genes become of interest for the community, they are simultaneously sequenced in several organisms. Waiting for all genes from all species to be sequenced before proceeding to an analysis is of course not an option. The current standard is to use as many genes as available: concatenating several genes into one supergene increases the sample size – here the gene length – and implies a more accurate analysis. Such concatenation implicitly assumes that every new gene has the same evolutionary history as the others. Unfortunately, there is no certainty about that. It is well known that many mechanisms – recombination, selective sweep,

purifying or positive selection among others (Balding et al., 2007)– lead different genes to have different histories. When a new gene becomes available, it should thus be tested for consistency before being included in the sample. If there is suspicion or exterior information that the new gene do not share a common history with the previous ones, the focus is on the minimum gene length necessary to confidently assess the difference, as in the optimization of the change point detection.

The issue of change point detection is hardly new but unlike most methods available in the sequential tests literature the alternative hypothesis is not well specified: a gene can be affected by a number of evolutionary event and thus have a number of evolutionary histories. Specifying one, or even a finite set, of those histories in H_1 is hardly better than an educated guess. The main focus is thus on rejecting H_0, close in philosophy to the Repeated Significance Test (RST) (Armitage et al., 1969; O'Brien and Fleming, 1979; Pocock, 1977). This particular issue of assessing consistency when the alternative is not well specified can also be found in the online learning literature and is there referred to as concept drift (Domingos and Hulten, 2000).

The article is organized as follows: Section B.2 introduces the key concepts and provides intuition about the kind of results we expect. Section B.3 present our main results, derived from Edgeworth expansions, and discuss their strong and weak points. Section B.4 builds upon the results of Section B.3 to present a test of consistency of a new set of data with previous ones. Proofs are postponed to Section B.5.

B.2 Definitions and Notations

B.2.1 Definition of $\Delta_{n,+k}$ and $\Delta_{n,-k}$

Let $(X_1, \ldots, X_n, \ldots)$ be a sequence of i.i.d random variables whose common distribution function is F_0. Consider the sample mean for the first n terms:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and define:

$$\begin{aligned} \Delta_{n,+k} &= \overline{X}_{n+k} - \overline{X}_n, \\ \Delta_{n,-k} &= \overline{X}_{n-k} - \overline{X}_n. \end{aligned}$$

Since $\Delta_{n,+k}$ is invariant by translation of the X_is, we assume without loss of generality that the X_k are centered ($E[X_1] = \mu = 0$) and furthermore note:

$$E[X_1^2] = \sigma^2 \quad E[X_1^3] = \kappa \quad E[|X_1|^3] = \beta_3 < \infty$$

A alternative definition of $\Delta_{n,+k}$ is

$$\Delta_{n,+k} = \frac{1}{n+k} \sum_{j=1}^{k} X_{n+j} - \frac{k}{n(n+k)} \sum_{j=1}^{n} X_j. \tag{B.1}$$

$\Delta_{n,+k}$ (resp. $\Delta_{n,-k}$) is centered with distribution function F_+ (resp. F_-) and variance $\sigma^2_{n,+k}$ (resp. $\sigma^2_{n,-k}$) where

$$\sigma^2_{n,+k} = \frac{k\sigma^2}{n(n+k)} \quad \text{and} \quad \sigma^2_{n,-k} = \frac{k\sigma^2}{n(n-k)}$$

$\Delta_{n,+k}$ (resp. $\Delta_{n,-k}$) represent perturbations of the sample mean induced by adding (resp. removing) k units from the sample. As one would expect, when n increases perturbations to the sample mean are the same no matter whether k terms are added to or removed from the sample. To formalize this intuition, we focus on the difference $F_+ - F_-$. $F_+(x) - F_-(x)$ is convenient for at least two results: using appropriate expansion techniques, we can get results about its order of magnitude and $\sup_{x\in\mathbb{R}} |F_+(x) - F_-(x)|$, the quantity of interest in Kolmogorov-Smirnoff test, is easy to calculate given some expansion of $F_+(x) - F_-(x)$.

B.2.2 Characteristic Function

But, before proceeding to derivation of the expansion, we recall a few properties of characteristic functions and use them to get insight into the difference between $\Delta_{n,+k}$ and $\Delta_{n,-k}$.

Let X be a real valued random variable with distribution function F_X. Let f_X be the characteristic function of X defined as $f_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} dF_X(x)$.

Hereafter and unless specified otherwise, we use the shorthands f for f_X, f_+ for $f_{\Delta_{n,+k}}$ and f_- for $f_{\Delta_{n,-k}}$. Thanks to Eq. (B.1) and classical properties of the characteristic function for independent random variables, we have

$$f_+(t) = f\left(\frac{t}{n+k}\right)^k f\left(\frac{-t}{n(n+k)}\right)^n. \tag{B.2}$$

Taylor expansion around 0 yields

$$f_-(t) - f_+(t) \simeq \frac{kt^2}{\sigma^2 n^2}\frac{k}{n}.$$

where lower order terms have been omitted. Note that $Var(\Delta_{n,+k}) \sim Var(\Delta_{n,-k}) \sim \frac{k\sigma^2}{n^2}$. Normalizing $\Delta_{n,+k}$ and $\Delta_{n,-k}$ so that they have asymptotic variance 1 and considering the difference between the characteristic function of the normalized version yields

$$f_-\left(\frac{nt}{\sqrt{k}\sigma}\right) - f_+\left(\frac{nt}{\sqrt{k}\sigma}\right) \simeq \frac{kt^2}{n}. \tag{B.3}$$

omitting again all lower order terms. Since the first order term in the expansion of $f_- - f_+$ around 0 is of order k/n and although local expansion provides is not enough to prove it, we expect from the inversion theorem the difference $F_- - F_+$ to be of order k/n. However, in order to achieve this result, two competing speeds need to be balanced: $k^{-1/2}$ and k/n. An intuitive justification follows. It is clear that

$$\frac{n}{\sqrt{k}\sigma}\Delta_{n,+k} = \left(1+\frac{k}{n}\right)^{-1} \frac{1}{\sigma\sqrt{k}} \sum_{j=1}^{k} (X_{n+j} - \bar{X}_n) \quad \text{and} \quad \frac{n}{\sqrt{k}\sigma}\Delta_{n,-k} = \frac{1}{\sigma\sqrt{k}} \sum_{j=1}^{k} (X_{n-k+j} - \bar{X}_{n-k}) \tag{B.4}$$

where $\bar{X}_n$ is the empirical mean of an n-sample of i.i.d. X_j. Since $\bar{X}_n = \mu + O_P\left(n^{-1/2}\right)$, it is clear from Eq. (B.4) that $\frac{n}{\sqrt{k}\sigma}\Delta_{n,+k}$ can be thought of as the standardized sum of k i.i.d roughly centered random variables with variance 1. If k goes to infinity with n, the speed $k^{-1/2}$ is thus the usual speed of the central limit theorem whereas k/n is the speed of the first order difference between variance of $\Delta_{n,+k}$ and $\Delta_{n,-k}$. Depending on the regularity of F and the compared speed of $k^{-1/2}$ and k/n, we can make the intuition rigorous and prove the assertion:

$$F_+\left(\frac{\sqrt{k}\sigma x}{n}\right) - F_-\left(\frac{\sqrt{k}\sigma x}{n}\right) = \frac{x}{\sqrt{2\pi}}e^{-x^2/2}\frac{k}{n} + o\left(\frac{k}{n}\right) \tag{B.5}$$

uniformly in x. Proper formulations and proofs are provided in Section B.3.

Eq. (B.3) provides an asymptotic expansion of $f_+ - f_-$ in an interval around 0 and, although it gives some insight about the resulting Eq. (B.5), it is not powerful enough to derive it properly. We therefore resort to Edgeworth expansion, with an Edgeworth series acting as a middleman between f_+ and f_-. This is the aim of Section B.3.

B.3 Edgeworth Expansion

Edgeworth series provide an approximation of a probability distribution in terms of its cumulants and are an improvement to the central limit theorem. The nice property of Edgeworth expansions is that they are true asymptotic expansions. We can thus control the error between a probability distribution and its Edgeworth expansion. The literature about Edgeworth expansion is quite abundant and full of powerful results. However most, if not all, of these results rely heavily on f satisfying the so-called *Cramér's Condition*:

$$\limsup_{|t|\to\infty} |f(t)| < 1 \tag{B.6}$$

Cramér's condition is equivalent to F having an absolutely continuous component (Hall, 1984) but we take a special interest in non-lattice completely discontinuous F (*i.e.* discrete X) for which condition (B.6) is not satisfied. We deal with distribution functions satisfying Cramér's condition in Section B.3.1 before turning to non-lattice discrete distribution functions in Section B.3.2. Proofs are postponed in Section B.5.

B.3.1 With Cramér's Condition

The main result of this section is the following:

Theorem 16 *Let (X_i) be a sequence of i.i.d. real valued random variables with distribution function F. Suppose that Cramér's condition holds,* i.e. *that* $\limsup_{|t|\to\infty} |f(t)| < 1$. *Suppose furthermore that there exists an integer $m \geq 1$ such $E[|X|^{m+2}] < \infty$ and consider $\alpha \in \left(\frac{2}{m+2}, 1\right)$. If $k \sim n^\alpha$ then:*

$$F_+\left(\frac{\sqrt{k}\sigma x}{n}\right) - F_-\left(\frac{\sqrt{k}\sigma x}{n}\right) = \frac{xe^{-x^2/2}}{\sqrt{2\pi}}\frac{k}{n} + o\left(\frac{k}{n}\right) \tag{B.7}$$

uniformly in x.

If $E[|X|^m] < \infty$ for all m, as is the case for gaussian random variables, α can take any value in $(0, 1)$. The only missing case is $k = o(n^\varepsilon)$ for all $\varepsilon > 0$. In particular and unlike gaussian variables, as will be shown in Prop. 18, k can not be fixed or grow only logarithmically with n.

B.3.2 Without Cramér's Condition

The main result of this section is the following:

Theorem 17 *Let (X_i) be a sequence of i.i.d. real valued random variables with distribution function F. Suppose that X is a non lattice, discrete random variable. Suppose furthermore that $\beta_3 = E[|X|^3] < \infty$ and consider $\alpha \in \left(\frac{2}{3}, 1\right)$. If $k \sim n^\alpha$ then:*

$$F_+\left(\frac{\sqrt{k}\sigma x}{n}\right) - F_-\left(\frac{\sqrt{k}\sigma x}{n}\right) = \frac{xe^{-x^2/2}}{\sqrt{2\pi}}\frac{k}{n} + o\left(\frac{k}{n}\right) \tag{B.8}$$

uniformly in x.

The fundamental difference between Theorems 16 and 15 lies in the range of value α can take. When the distribution function F of X has some absolutely continuous component, k is allowed, upon moment conditions, to grow slowly compared to n. When the distribution function is completely discrete, the third order moment is enough to achieve the expansion. Higher order moments, even if they do exist, are not sufficient to expand the range of value α can take and are thus not required.

B.3.3 New Generating Process

The main result of this section is the following:

Theorem 18 *Let X_i be a sequence of independent real valued random variables such with distribution function F_0 for $i \leq n$ and distribution function F_1 for $i > n$. Let $\Delta_{n,+k} = \overline{X}_{n+k} - \overline{X}_n$ and F_+ its distribution function. Suppose that F_0 (resp. F_1) has finite expectation μ_0 (resp. μ_1) and variance σ_0^2 (resp. σ_1^2). Suppose furthermore that $\beta_3 = E[|X_{n+1}|^3] < \infty$ and consider $\alpha \in (0, 1)$. If $k \sim n^\alpha$, then:*

$$F_+\left(\frac{\sqrt{k}\sigma_1 x}{n}\right) = \Phi\left(x - \frac{n}{n+k}\frac{\sqrt{k}(\mu_1 - \mu_0)}{\sigma_1}\right) + O(n^{-\beta}) \tag{B.9}$$

uniformly in x, where $\beta = \min\left(\frac{\alpha}{2}, 1 - \alpha\right)$. If x is restricted to a bounded range and $\mu_1 \neq \mu_0$, the correcting term $n/(n+k)$ is unnecessary and Eq. (B.9) *simplifies to*

$$F_+\left(\frac{\sqrt{k}\sigma_1 x}{n}\right) = \Phi\left(x - \frac{\sqrt{k}(\mu_1 - \mu_0)}{\sigma_1}\right) + O(n^{-\beta}). \tag{B.10}$$

Theorem 17 requires a third order condition on the new generating process Y to ensure that the remaining term is of order $O(k^{-1/2})$. Neglecting second order terms, $\frac{n\Delta_{n,+k}}{\sqrt{k}\sigma_1}$ behaves like a gaussian variable with mean $\sqrt{k}\frac{\mu_1-\mu_0}{\sigma_1}$ and variance 1. As we could expect, the mean diverges faster if μ_0 and μ_1 are well separated when compared to the scale σ_1.

B.3.4 About Discrete Distributions

Our motivating example of DNA analysis is intimately linked to discrete state space. When comparing the same gene among a set of s organisms, each nucleotide in a species is associated to its homologous in the remaining species. An observation consists of a s-uple of nucleotides,. Each nucleotide can take value in the set $\{A, C, G, T\}$ and thus the s-uples take value in $\{A, C, G, T\}^s$. The statistic of interest is the likelihood of an observation under a given model. The observations are intrinsically discrete and so is the likelihood of an observation under a given model. To turn these likelihoods to continuous variables and allow for the use of Theorem 15 instead of the less powerful Theorem 16, we must resort to the trick exposed hereafter.

Formally, consider a discrete space $A = (a_i)_{i=1,\dots,N}$ and a probability measure $\theta = (\theta_1, \dots, \theta_N)$ on A. In DNA analysis, $A = \{A, C, G, T\}^s$ and θ is a model assigning a probability to each $a \in A$. Assume $\theta_i > 0$ for all i and let $(Z_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables such that $P(Z = a_j) = \theta_j$ for $j = 1, \dots, N$. We take a special interest in $(X_i)_{i\in\mathbb{N}}$ defined as

$$X_i = \log P(\{Z_i\}) = \sum_{j=1}^{N} \log P(Z_i = a_j)\mathbb{1}_{\{Z_i=a_j\}}$$

(X_i) is easily an i.i.d sequence of discrete random variables such that $P(X = \log(\theta_j)) = c_j\theta_j$ where c_j is the number of outcomes a_k such that $\theta_k = \theta_j$. For Theorem 16 to apply here, X should be non-lattice. The only way X can be lattice is if there exists some $0 < u < 1$ and some $v \in \mathbb{R}$ such that $\log_u \theta_j \in v + u\mathbb{Z}$ for all $1 \le j \le N$. This include truncated geometric distribution and Bernoulli distribution but not for example binomial (with $n \ge 2$). Under non-latticeness, we can prove thanks to Theorem 16 that $\sup_{\mathbb{R}} |F_+ - F_-| = \frac{1}{\sqrt{2\pi e}}\frac{k}{n} + o\left(\frac{k}{n}\right)$ but only if $k \sim n^\alpha$ with $\alpha \in (2/3, 1)$. We don't have access to lower values of α.

Suppose now that θ is not the same for all Z_i but rather that each Z_i is drawn from A according to a specific $\alpha^{(i)} = (\alpha_1^{(i)}, \dots, \alpha_N^{(i)})$ and furthermore that $\alpha^{(\mathbf{i})}$ is an i.i.d sequence from a Dirichlet distribution $\mathrm{Dir}(\lambda\theta)$ that has density:

$$f(v_1, \dots, v_{N-1}) = \frac{\prod_{i=1}^{N}\Gamma(\lambda\theta_i)}{\Gamma(\lambda\sum_{i=1}^{N}\theta_i)}\prod_{i=1}^{N-1} v_i^{\lambda\theta_i - 1}$$

for all $v_1, \dots, v_{N-1} > 0$ such that $\sum_{i=1}^{N-1} v_i < 1$ and $V_N = 1 - \sum_{i=1}^{N-1} V_i$. Intuitively, $(V_1, \dots, V_N)$ is a vector of the N dimensional unit simplex with mean θ and variance inversely proportional to λ: the marginal distribution of V_i has mean θ_i and variance $\frac{\theta_i(1-\theta_i)}{\lambda+1}$. Using $Dir(\lambda\theta)$ instead of θ can be seen as a regularization of the previous case, with θ being the limiting case of $Dir(\lambda\theta)$ when λ goes to infinity.

It is then easily seen that the X_i are i.i.d random variables taking value in $\mathbb{R}_-$ and absolutely continuous with respect to the Lebesgue-measure. A bit of algebra gives for all m

$$\begin{aligned} E[|X|^m] &= E\left[\sum_{i=1}^{N} |log^m P(Z=a_i)| \mathbb{1}_{Z=a_i}\right] = \sum_{i=1}^{N} \int_0^1 |\log(\alpha_i)|^m \alpha_i p(\alpha_i|\theta) d\alpha_i \\ &= \sum_{i=1}^{N} \frac{\Gamma(\lambda\theta_i)\Gamma(\lambda(1-\theta_i))}{\Gamma(\lambda)} \int_0^1 |log^m(x)| x^{\lambda\theta_i}(1-x)^{\lambda(1-\theta_i)-1} dx \\ &< \infty \end{aligned}$$

In this case of particular interest, Theorem 15 applies for any value of α in $(0,1)$ as m can be taken arbitrary large.

B.4 Application to Test

Theorems 15 and 16 are useful for detecting changes in the generating process of new observations.

We want to test whether the new batch of observations is generated by the same process as the previous observations. Formally, given two probability distributions F_0 and F_1, and a sequence of independent random variables (X_i) with associated distribution function F_{X_i}, we want to test H_0: "$F_{X_i} = F_0$ for $i = 1, \ldots, n+k$" against H_1: "$F_{X_i} = F_0$ for $i \leq n$ and $F_{X_i} = F_1$ otherwise".

In our problem, the statistic of interest is the sample mean, calculated either on all $n+k$ observations ($\overline{X}_{n+k}$)or only the previous n observations ($\overline{X}_n$). We shall therefore assume that F_0 and F_1 have different means μ_0 and μ_1 and assess model shift by shift in the sample mean. $\Delta_{n,+k} = \overline{X}_{n+k} - \overline{X}_n$ represents the influence of the batch of k new observations on the mean, *i.e* the translation of the sample mean induced by adding the batch of new observation to the calculation. The use of the term "influence" is not coincidental: $\Delta_{n,+k}$ is strongly connected to influence functions (Hampel, 1974a; Huber, 2004). When the quantity to estimate is the mean μ of a distribution and $k = 1$, $n\Delta_{n,+1}$ is indeed exactly the empirical influence value of observation X_{n+1} on the estimator $\hat{\mu} = \overline{X}_n$ of μ, *i.e.* the influence of an infinitesimal perturbation on $\hat{\mu}$ along the direction δ_{X_i}, the unit mass at point X_i.

Large positive or negative influence values point up the corresponding observations as potentials outliers whereas small to moderate influence values support consistency of the data. Up to a rescaling, $\Delta_{n,+k}$ can be understood as an extension of influence functions to a batch of observations instead of a single one.

B.4.1 Distribution of $\Delta_{n,+k}$ under H_0

Let $k \in \{n^{\beta_1}, n^{\beta_2}\}$ with β_1 and β_2 to be specified later. Under H_0, $F_{X_i} = F_0$ for $i = 1, \ldots, n+k$ and it comes from Theorems 15 for continuous and 16 for discrete distributions that $\Delta_{n,+k}$ and $\Delta_{n,-k}$ have the same distribution function, up to a correcting

term of order k/n. For discrete distributions, $(\beta_1, \beta_2) = (2/3 + \varepsilon, 1 - \varepsilon)$ where ε is an arbitrary small positive value. For continuous distributions $(\beta_1, \beta_2) = (\frac{2}{m+2} + \varepsilon, 1 - \varepsilon)$ where ε is again an arbitrary small positive value and m is the highest order moment of F_0.

The alternative definition Eq. (B.1) of $\Delta_{n,-k}$ gives different weights to $(X_1, \ldots, X_{n-k})$ and $(X_{n-k+1}, \ldots, X_n)$. Under H_0, the first n observations are identically distributed and exchangeable. Exchangeability implies that the order of $(X_1, \ldots, X_n)$ does not matter. Since their order does not matter, $(X_{n-k+1}, \ldots, X_n)$ can be replaced by any other subset of $(X_1, \ldots, X_n)$ of size k. In particular, the distribution of $\Delta_{n,-k}$ can be approximated by repeatedly selecting k terms from $(X_1, \ldots, X_n)$ and substituting them to $(X_{n-k+1}, \ldots, X_n)$.

When the distribution F_0 of the X_i under H_0 is not a simple parametric function or involves a large number of parameters, the exact distribution function of $\Delta_{n,+k}$ is unachievable. Even an Edgeworth expansion *à la* Prop. 27 requires the estimation of many cumulants. By contrast a good numerical approximation of F_- is available thanks to the previous remark and we can substitute it to F_+. Adding the correcting term of order k/n only requires the estimation of the standard deviation σ of F_0. And one may notice that since there are $n + k$ observations with n larger than k, the estimation of σ is significantly more accurate than the approximation of F_- by its empirical version.

Wrapping up the preceding remarks, the distribution F_+ of $\Delta_{n,-k}$ can approximated in the following way:

(i) Compute the mean $\overline{X}_n$ of the n observations;

(ii) Select at random without replacement k observations among the n;

(iii) Compute the mean $\overline{X}^\star_{n-k}$ of the remaining $n - k$ observations;

(iv) Record the difference $\Delta^\star_{n,-k} = \overline{X}_n - \overline{X}^\star_{n-k}$;

(v) Repeat (ii) to (iv) a large number (N) of times.

The distribution F_+ of $\Delta_{n,+k}$ is then well approximated by the distribution of $\Delta^\star_{n,-k}$, corrected by the term of order k/n (see Hall (1984) for more detailed results). The approximation of F_+ can then be used to construct a critical region for rejecting H_0 based on the $\Delta_{n,+k}$.

B.4.2 Distribution of $\Delta_{n,+k}$ under H_1

Under H_1, noting σ_1^2 the variance of the distribution F_1 and assuming $\mu_0 \neq \mu_1$, Theorem 17 implies

$$F_+\left(\frac{\sqrt{k}\sigma_1 t}{n}\right) = \Phi\left(t - \sqrt{k}\frac{\mu_1 - \mu_0}{\sigma_1}\right) + O(k^{-1/2}) + O\left(\frac{k}{n}\right)$$

where Φ is the standard normal distribution. The distribution of $\Delta_{n,+k}$ under H_1 is approximately gaussian with mean $\sqrt{k}\frac{\mu_1-\mu_0}{\sigma_1}$ diverging to ∞ with k. Difference between F_+ and F_- is of order $O(1)$ and terms correcting for the lack of gaussianity of the observations are negligible in front of the main term. Given the boundary of the rejection zone calculated in section B.4.1, the approximate power of the test can then easily be computed.

B.4.3 Calibrating the test

To test H_0: "$F_{X_i} = F_0$ for $i = 1, \dots, n+k$" against H_1: "$F_{X_i} = F_0$ for $i \leq n$ and $F_{X_i} = F_1$ otherwise" at the level α when F_0 and F_1 have different expectations μ_0 and μ_1, we adopt the test statistic $\Delta_{n,+k}$.

$\Delta_{n,+k}$ is centered under H_0 but not under H_1, the rejection zone is thus of the form $\{\Delta_{n,+k} \leq A_{\alpha/2}\} \cup \{\Delta_{n,+k} \leq A_{1-\alpha/2}\}$. $A_{\alpha/2}$ and $A_{1-\alpha/2}$ would ideally be the $\alpha/2$ and $1-\alpha/2$ quantiles of $\Delta_{n,+k}$ under H_0. Unfortunately, the distribution of $\Delta_{n,+k}$ maybe be unknown (if F_0 is unknown) or untractable (if F_0 is a discrete distribution with a large number of outcomes). One way to avoid this problem would be to estimate them by bootstrap. The drawback is that bootstrap estimation is relevant only under H_0; under H_1, the quantiles would not be correctly estimated and the test would be unproperly calibrated.

However, replacing $A_{\alpha/2}$ and $A_{1-\alpha/2}$ by $B_{\alpha/2}$ and $B_{1-\alpha/2}$, the $\alpha/2$ and $1-\alpha/2$ quantiles of $\Delta_{n,-k}$ yields a error of order $2k/n$ on the level of the test. The distribution function of $\Delta_{n,-k}$ is no better known than the distribution function of $\Delta_{n,+k}$ but unlike $A_{\alpha/2}$ and $A_{1-\alpha/2}$, the quantiles $B_{\alpha/2}$ and $B_{1-\alpha/2}$ can be correctly estimated by bootstrap *both* under H_0 and H_1.

B.4.4 Discussion of the results

About the remainder term: Theorems 15 and 17 are derived for very general distribution functions: they hold under mere moment conditions. When the distribution at hand is better specified, more accurate results can reasonably be expected. But in the absence of any further assumptions, the remainder of order $o(k/n)$ is possibly the best we can achieve.

For example, if the distribution function is skewed, tedious calculations show that the remainder is at least of order $O(\sqrt{k}/n)$. And we can get closer to k/n by mimicking discrete lattice distributions. Lattice distributions are off-limits but can be seen as the limiting case of non-lattice discrete distributions: a discrete non-lattice distribution with jumps of size $1/2 - \varepsilon$ at points ± 1 and size ε at points $\pm\sqrt{2}$ is very close to a lattice distribution with jumps of size $1/2$ at points ± 1 for small enough ε. For the limiting case of F_0 being such a lattice distribution, and for odd k such that neither n/k nor $(n-k)/k$ are integer, F_+ has a jump of size of asymptotic size $\sqrt{2/\pi k}$ at point $1/(n+k)$ when F_- has no jump at that point. Since $\frac{kx}{n}e^{-x^2/2}$ has no jump whatsoever at any point, the extremum of $(F_+(\sqrt{k}\sigma x/n) - F_-(\sqrt{k}\sigma x/n)) - kx/ne^{-\frac{x^2}{2}}$ is at least $\sqrt{2/\pi k}$ attained for $x = \frac{n}{n+k}\frac{1}{\sqrt{k}\sigma}$ and thus of order at least $k^{-1/2}$. Since $k^{-1/2} \sim n^{-\alpha/2}$ which can be arbitrarily close to k/n as α decreases towards $2/3$, the $o(k/n)$ can not be improved upon in this case.

On the other hand, gaussian variables have such a nice distribution that most calculations about F_- and F_+ can be done exactly. Most important of all, whatever the value of k, if $(X_{n+1}, \dots, X_{n+k})$ is a linear vector, then any linear combination of $X_{n+1}, \dots, X_{n+k}$ is gaussian. Going back to Eq. (B.4), the first term is *exactly* gaussian and there is no need whatsoever for correcting terms of order $k^{j/2}$. This is the most

favorable case, for which the remainder in Theorem 15 has the smallest order of magnitude.

Under H_0, if the X_i have mean μ and variance σ^2, then $\frac{n\Delta_{n,-k}}{\sqrt{k}\sigma} \sim \mathcal{N}(0, \frac{n}{n-k})$, $\frac{n\Delta_{n,+k}}{\sqrt{k}\sigma} \sim \mathcal{N}(0, \frac{n}{n+k})$ and we can derive the following result:

Proposition 19 *Let $\Delta_{n,+k}$ and $\Delta_{n,-k}$ be defined as before, then:*

$$F_+\left(\frac{\sqrt{k}\sigma x}{n}\right) - F_-\left(\frac{\sqrt{k}\sigma x}{n}\right) - \frac{k}{n}\frac{xe^{-\frac{x^2}{2}}}{\sqrt{2\pi}} = O\left(\frac{k^3}{n^3}\right)$$

Uniformly in x.

Prop 18 is better than the result provided by Theorem 15, as $O(k^3/n^3)$ is smaller than $o(k/n)$. Further algebra can even prove here that $O(k^3/n^3)$ is no greater than $1.2k^3/n^3$, uniformly in x.

Under hypothesis H_1, we have:

$$\frac{n\Delta_{n,+k}}{\sqrt{k}\sigma} \sim \mathcal{N}\left(\frac{n}{n+k}\frac{\sqrt{k}(\mu_1-\mu_0)}{\sigma_1}, 1 + \frac{Ak}{n}\right)$$

where $A \leq 1 + \frac{\sigma_0^2}{\sigma_1^2}$. As expected, the result is again slightly more accurate than would be obtained by Theorem 17 alone, as the remainder is exactly, instead of at least, of order $(k/n)^{1/2}$. In the gaussian case, we can thus easily improve upon results from Section B.3.

About Cramér's Condition: Cramér's condition plays a crucial role in the demonstration of Theorem 15. Without Cramér's condition, there is no guarantee that jumps of the distribution function F_+ are of order $o(k^{-1})$ and higher order moments of F_+ can not be used to improve the range of k that can be used. Indeed, as the binomial example emphasizes for the forbidden but limiting case of lattice distribution, jumps can be of order $k^{-1/2}$. But for non-lattice discrete lattice distributions, the maximum jump is at most of order $o(k^{-1/2})$ and can be much smaller than that, for example $o(k^{-1})$. In this case, is might be possible upon further work to increase the range of value α can take in Theorem 16.

Remaining Work: There are two main caveats to our test. The first one concerns the asymptotic error rate we make when constructing the rejection zone. We know from Sec. **??** that the total error comes first from replacing the distribution F_+ by F_- and then from replacing F_- by its bootstrap counterpart $F_-^\star$. The first error rate is of order k/n and can reduced be to order $o(k/n)$. The second one is not studied here but its order of magnitude should be compared with k/n. The second caveat comes from the application to phylogenetics. In a phylogenetic framework, the X_i would be the likelihood score of a site Z_i, as discussed in Sec. **??**. In the current state, the test suppose that the distribution F_0 of the Z_i is known so that the X_i can be properly calculated. However F_0 is unknown in general. Replacing F_0 with $\widehat{F_0}$ would be more realistic, but doing this introduces an additional error term in the the test which should also be compared to k/n.

B.5 Proofs

Before we proceed to proof of Theorem 15, 16 and 17, we recall some lemma concerning the expansion of $f^k(x/\sqrt{k})$.

Without loss of generality, we assume $E[X] = 0$. Note σ^2 the variance of X, $\alpha_j = E[X^j]$ the moment of order j and κ_j the j-th cumulant of X, defined as:

$$\kappa_j = \frac{1}{i^j}\frac{\mathrm{d}^j}{\mathrm{d}t^j}\ln E\left[e^{itX}\right]\Big|_{t=0} = \frac{1}{i^j}(\ln \circ f)^{(j)}(0)$$

B.5.1 Previous Results

Lemma 20 (Esseen45) *Let (X_i) a sequence of i.i.d. random variables and $m \geq 3$ an integer such that $E[|X|^m] < \infty$, then*

$$\left|f_X\left(\frac{t}{\sqrt{k}\sigma}\right)^k - e^{-\frac{t^2}{2}}\left(1 + \sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}\right)\right| \leq \frac{\delta(k)}{k^{\frac{m-2}{2}}}(|t|^m + |t|^{3(m-1)})e^{-\frac{t^2}{4}} \quad \textit{for} \quad |t| \leq \frac{\sigma\sqrt{k}}{4\beta_m^{1/m}}$$

where $P_j(it) = \sum_{v=1}^{j} c_{jv}(it)^{2v+j}$ is a polynomial of degree $3j$ in it, the coefficient c_{jv} being a polynomial in the cumulants $\kappa_3, \ldots, \kappa_{j-v+3}$ and $\delta(k) \to 0$.

Lemma 21 (Esseen45) *Let (X_i) a sequence of i.i.d. random variables and $2 < \nu \leq 3$ a real number such that $\beta_\nu = E[|X|^\nu] < \infty$, then there exists a constant C_ν depending only on ν such that*

$$\left|f_X\left(\frac{t}{\sqrt{k}\sigma}\right)^k - e^{-\frac{t^2}{2}}\right| \leq \frac{C_\nu}{n^{\frac{\nu-2}{2}}}\frac{\beta_\nu}{\sigma^\nu}|t|^\nu e^{-\frac{t^2}{4}} \quad \textit{for} \quad |t| \leq \frac{\sigma^{\frac{1}{\nu-2}}\sqrt{k}}{(24\beta_\nu)^{\frac{1}{\nu-2}}}$$

Lemma 19 and 20 are proved in Esseen (1945) (p. 44). An alternative proof can be found in Cramer (1937) (p. 71 and 74).

Lemma 22 (Esseen48) *Let X be a non lattice discrete random variable, then for every $\eta > 0$ there exists a positive function $\lambda(k) \underset{k\to\infty}{\longrightarrow} \infty$ such that:*

$$\int_\eta^{\lambda(k)} \frac{|f(t)|^k}{t} = o\left(\frac{1}{\sqrt{k}}\right)$$

The proof of Lemma 20 can be found in Esseen (1945) (Lemma 1, p. 49).

We recall one last theorem before proceeding to the proof.

Theorem 23 (Essen48) *Let A, T and ε be arbitrary positive constants, $F(x)$ a non-decreasing function, $G(x)$ a real function of bounded variation on the real axis, $f(t)$ and $g(t)$ the corresponding Fourier-Stieltjed transforms such that:*

1. $F(-\infty) = G(-\infty) = 0$, $F(\infty) = G(\infty)$

2. $G'(x)$ *exists everywhere and* $|G'(x)| \leq A$
3. $\int_{-T}^{T} \left| \frac{f(t)-g(t)}{t} \right| dt = \varepsilon$

To every number $k > 1$*, there corresponds a finite positive number* $c(k)$*, only depending on* k*, such that*

$$|F(x) - G(x)| \leq k\frac{\varepsilon}{2\pi} + c(k)\frac{A}{T}$$

The proof of Theorem 22 is given in Esseen (1945) (Theorem 2.a, p. 32)

B.5.2 New Results

Lemma 23 is a generalization of Lemma 19.

Lemma 24 *Suppose that* X_i *is a sequence of i.i.d. random variables such* $E[|X|^m] < \infty$ *for an integer* $m \geq 3$*, then for* $|t| \leq \frac{\sigma\sqrt{k}}{4\beta_m^{1/m}}$*:*

$$\left| f_X\left(\frac{t}{\sqrt{k}\sigma}\frac{n}{n+k}\right)^k - e^{-\frac{t^2}{2}}\left(1+\frac{kt^2}{n}\right)\left(1+\sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}\right)\right| \leq \left\{\frac{\delta(k)}{k^{\frac{m-2}{2}}} + C_m\frac{\sqrt{k}}{n}\right\}(|t|^3 + |t|^{3(m-1)})e^{-\frac{t^2}{4}} + C'_m\frac{k^2}{n^2}(|t|^4 + |t|^{3m-2})e^{-\frac{t^2}{4}}$$

where $P_j(it) = \sum_{v=1}^{j} c_{jv}(it)^{2v+j}$ *is a polynomial of degree* $3j$ *in it, the coefficient* c_{jv} *being a polynomial in the cumulants* $\kappa_3, \ldots, \kappa_{j-v+3}$*,* $\lim_{k\to\infty}\delta(k) = 0$ *and* C_m *and* C'_m *are constants depending only on* m*.*

Proof. It follows from Lemma 19 that

$$\left| f_X\left(\frac{t}{\sqrt{k}\sigma}\frac{n}{n+k}\right)^k - e^{-\frac{t^2}{2}(1+\frac{k}{n})^{-2}}\left(1+\sum_{j=1}^{m-2}\frac{P_j(it(1+\frac{k}{n}))}{k^{j/2}}\right)\right| \leq \frac{\delta(k)}{k^{\frac{m-2}{2}}}(|t|^m + |t|^{3(m-1)})\left(1+\frac{k}{n}\right)^{-3(m-1)} e^{-\frac{t^2(1+\frac{k}{n})^{-2}}{4}}$$

We now expand $e^{-\frac{t^2}{2}(1+\frac{k}{n})^{-2}}$ in power of $\frac{k}{n}$ and arrange the terms in a convenient order.

$$-\frac{t^2}{2}\left\{\left(1+\frac{k}{n}\right)^{-2} - 1\right\} = \frac{kt^2}{n} - \frac{k^2t^2}{2(n+k)^2}$$

Furthermore $-\frac{t^2}{2} \leq -\frac{t^2}{2}\left(1+\frac{k}{n}\right)^{-2} \leq -\frac{t^2}{4}$, where the last inequality holds for large enough n. It then follows from a Taylor expansion that

$$\left| e^{-\frac{t^2}{2}(1+\frac{k}{n})^{-2}} - e^{-\frac{t^2}{2}}\left(1+\frac{kt^2}{n}\right)\right| \leq e^{-\frac{t^2}{2}(1+\frac{k}{n})^{-2}}\left(\frac{kt^2}{n} - \frac{k^2t^2}{2(n+k)^2}\right)^2 \leq e^{-\frac{t^2}{4}}\frac{k^2t^4}{n^2}.$$

We also have, for any integer j

$$(it)^j\left(1+\frac{k}{n}\right)^{-j}-(it)^j=-(it)^j\left\{\frac{jk}{n}+O\left(\frac{k^2}{n^2}\right)\right\}.$$

And thus there exist a constant K_j, not depending on n and j such that

$$\left|P_j(it(1+\frac{k}{n}))-P_j(it)\right|=\left|\sum_{v=1}^{j}c_{jv}(it)^{2v+j}\left\{\frac{(2v+j)k}{n}+O\left(\frac{k^2}{n^2}\right)\right\}\right|\leq K_j(|t|^{j+2}+|t|^{3j})\frac{k}{n}$$

It follows that there exists a positive constant C_m, depending neither on n nor k such that

$$\left|1+\sum_{j=1}^{m-2}\frac{P_j(it(1+\frac{k}{n}))}{k^{j/2}}-\left(1+\sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}\right)\right|\leq\sum_{j=1}^{m-2}K_j\frac{k}{n}\frac{(|t|^{j+2}+|t|^{3j})}{k^{j/2}}\leq\frac{C_m}{3}\frac{\sqrt{k}}{n}(|t|^3+|t|^{3(m-2)}).$$

Finally $e^{-\frac{t^2}{2}}\left(1+\frac{kt^2}{n}\right)\leq 3e^{-\frac{t^2}{4}}$ and there exists a constant C'_m such that $\left|1+\sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}\right|\leq C'_m(1+|t|^{3(m-2)})$. For any four reals A,B,a,b, $|AB-ab|\leq|A(B-b)|+|b(A-a)|$. Using $A=e^{-\frac{t^2}{2}}\left(1+\frac{kt^2}{n}\right)$, $a=e^{-\frac{t^2}{2}(1+\frac{k}{n})^{-2}}$, $B=1+\sum_{j=1}^{m-2}\frac{P_j(it(1+\frac{k}{n}))}{k^{j/2}}$ and $b=1+\sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}$ we obtain:

$$\left|e^{-\frac{t^2}{2}(1+\frac{k}{n})^{-2}}\left(1+\sum_{j=1}^{m-2}\frac{P_j(it(1+\frac{k}{n}))}{k^{j/2}}\right)-e^{-\frac{t^2}{2}}\left(1+\frac{kt^2}{n}\right)\left(1+\sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}\right)\right|\leq$$
$$C_m\frac{\sqrt{k}}{n}(|t|^3+|t|^{3(m-2)})e^{-\frac{t^2}{4}}+C'_m\frac{k^2}{n^2}(|t|^4+|t|^{3m-2})e^{-\frac{t^2}{4}}$$

From which the result immediately follows. ■

Lemma 25 *With the notations previously defined and under the conditions of Theorem 15*

$$\left|f_X\left(\frac{-\sqrt{k}t}{(n-k)\sigma}\right)^{n-k}-\left(1-\frac{k}{n}\frac{t^2}{2}\right)\right|\leq K_-(t^2+t^4)\frac{k^2}{n^2}$$
$$\left|f_X\left(\frac{-\sqrt{k}t}{(n+k)\sigma}\right)^{n}-\left(1-\frac{k}{n}\frac{t^2}{2}\right)\right|\leq K_+(t^2+t^4)\frac{k^2}{n^2}$$

uniformly for $|t|\leq\frac{\sigma\sqrt{k}}{4\beta_m^{1/m}}$, where K_+ and K_- are constants not depending on n,k or X.

Proof. Since the two inequalities are proved in the same way, we prove only the first one. It is readily observed that $\beta_m^{1/m}$ increases with m, thus $\beta_3\leq\beta_m^{3/m}$. It follows by taking $\nu=3$ in Lemma 20 that for $|t|\leq\frac{\sigma\sqrt{k}}{4\beta_3^{1/3}}$,

$$\left|f_X\left(\frac{t}{\sqrt{k}\sigma}\right)^k-e^{-\frac{t^2}{2}}\right|\leq C_3\frac{\beta_3}{\sigma^3}\frac{1}{k^{1/2}}|t|^3e^{-\frac{t^2}{4}}$$

A simple decomposition of the quantity to upper bound yields

$$\left| f_X\left(\frac{-\sqrt{k}t}{(n-k)\sigma}\right)^{n-k} - \left(1-\frac{k}{n}\frac{t^2}{2}\right)\right| \leq \left| f_X\left(\frac{-\sqrt{k}t}{(n-k)\sigma}\right)^{n-k} - \exp\left\{-\frac{kt^2}{2(n-k)}\right\}\right| + \left|\exp\left\{-\frac{kt^2}{2(n-k)}\frac{t^2}{2}\right\} - \exp\left\{-\frac{kt^2}{2n}\right\}\right| + \left|\exp\left\{-\frac{kt^2}{2n}\right\} - \left(1-\frac{kt^2}{2n}\right)\right| \tag{B.11}$$

For large enough n, $k \leq n-k$ and thus for $|t| \leq \frac{\sigma\sqrt{k}}{4\beta_3^{1/3}} \leq \frac{\sigma\sqrt{n-k}}{4\beta_3^{1/3}}$, the first term of the right-hand side of Eq. (B.11) is upper bounded by

$$\left| f_X\left(\frac{-\sqrt{k}t}{(n-k)\sigma}\right)^{n-k} - e^{-\frac{k}{n-k}\frac{t^2}{2}}\right| \leq C_3\frac{\beta_3}{\sigma^3}\frac{k^{3/2}}{(n-k)^2}|t|^3 e^{-\frac{k}{n-k}\frac{t^2}{4}} \leq K_2\frac{k^{3/2}}{n^2}|t|^3$$

where $K_2 = C_3\beta_3/\sigma^3 \sup_n\{n^2/(n-k)^2\}$.

Using the classical inequality $|e^{x+y} - e^x| \leq |y|e^x$ for $y < 0$ we bound the second term of Eq. B.11:

$$\left|e^{-\frac{k}{n-k}\frac{t^2}{2}} - e^{-\frac{k}{n}\frac{t^2}{2}}\right| \leq e^{-\frac{k}{n}\frac{t^2}{2}}\left|\frac{k}{n-k} - \frac{k}{n}\right|\frac{t^2}{2} \leq K_1\frac{k^2}{n^2}t^2$$

where $K_1 = sup_n\{n/(n-k)\}/2$. Finally we bound the third term of Eq. B.11 using the inequality $|e^{-x} - (1-x)| \leq x^2/2$ for $x \geq 0$:

$$\left|e^{-\frac{k}{n}\frac{t^2}{2}} - \left(1-\frac{k}{n}\frac{t^2}{2}\right)\right| \leq \frac{k^2}{n^2}\frac{t^4}{4}$$

Since $k^{3/2}/n^2 = o(k^2/n^2)$, for K_+ large enough

$$K_2\frac{k^{3/2}}{n^2}|t|^3 + K_1\frac{k^2}{n^2}t^2 + \frac{1}{4}\frac{k^2}{n^2}t^4 \leq K_+(t^2+t^4)\frac{k^2}{n^2}$$

which ends the proof of the first part of the lemma. Replacing $n-k$ by $n+k$, the same demonstration holds and yields the second inequality of the lemma. ■

Lemma 26 *With the notations previously defined and under the conditions of Theorem 15*

$$\left| f_-\left(\frac{nt}{\sqrt{k}\sigma}\right) - e^{-\frac{t^2}{2}}\left(1-\frac{kt^2}{2n}\right)\left(1+\sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}\right)\right| \leq \frac{\delta(k)}{k^{\frac{m-2}{2}}}(|t|^3+|t|^{3(m-1)})e^{-\frac{t^2}{4}} + K'_-\frac{k^2}{n^2}e^{-\frac{t^2}{4}}(|t|^2+|t|^{3m-2})$$

$$\left| f_+\left(\frac{nt}{\sqrt{k}\sigma}\right) - e^{-\frac{t^2}{2}}\left(1+\frac{kt^2}{2n}\right)\left(1+\sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}\right)\right| \leq \left\{\frac{\delta(k)}{k^{\frac{m-2}{2}}} + C_m\frac{\sqrt{k}}{n}\right\}(|t|^3+|t|^{3(m-1)})e^{-\frac{t^2}{4}} + K'_+\frac{k^2}{n^2}e^{-\frac{t^2}{4}}(|t|^2+|t|^{3m-2})$$

uniformly for $|t| \leq \frac{\sigma\sqrt{k}}{4\beta_m^{1/m}}$, *where* K'_- *and* K'_+ *are constants not depending on n and k.*

Proof. For any four reals A, B, a, b, $|AB - ab| \leq |B(A-a)| + |a(B-b)|$. We take $A = f_X\left(\frac{t}{\sqrt{k}\sigma}\right)^k$, $a = e^{-\frac{t^2}{2}}\left(1 + \sum_{j=1}^{m-2} \frac{P_j(it)}{k^{j/2}}\right)$, $B = f_X\left(\frac{-t}{\sqrt{k(n-k)}\sigma}\right)^{n-k}$, $b = \left(1 - \frac{kt^2}{2n}\right)$. Using $|a| \leq C_m e^{-\frac{t^2}{2}}(1+|t|^{3(m-2)})$ and Lemma 24,

$$|a(B-b)| \leq C_m K_-(1+|t|^{3(m-2)})(t^2+t^4)e^{-\frac{t^2}{2}} \leq K'_- \frac{k^2}{n^2}(|t|^2+|t|^{3m-2})e^{-\frac{t^2}{4}}$$

where $K'_- = C_m K_- \sup_t \left\{ e^{-\frac{t^2}{4}} \frac{|t|^2+|t|^4+|t|^{3m-4}+|t|^{3m-2}}{|t|^2+|t|^{3m-2}} \right\}$. Similarly using $|B| \leq 1$ and Lemma 19

$$|B(A-a)| \leq \frac{\delta(k)}{k^{\frac{m-2}{2}}}(|t|^m + |t|^{3(m-1)})e^{-\frac{t^2}{4}}$$

Combining these two inequalities gives the result for the first part of the lemma. The second part is proved in the same way using Lemma 23 instead of 19. ■

B.5.3 Proof of Prop 18

Lemma 27 *Let Φ_a (resp. Φ_b) be the cumulative distribution function of a centered normal random variable with variance a (resp. b). Furthermore assume there is $\varepsilon > 0$ such that $a = (1+\varepsilon)^{-1}$ and $b = (1-\varepsilon)^{-1}$. Then, for vanishing ε:*

$$\Phi_a(x) - \Phi_b(x) = \varepsilon \frac{x e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} + O\left(\varepsilon^3\right)$$

uniformly in x.

Proof. Since $\Phi_{\sigma^2}(x) = \Phi(x/\sigma)$, we have $\Phi_a(x) = \Phi(x/\sqrt{a})$. By hypothesis $a^{-1/2} = (1+\varepsilon)^{1/2} = 1 + \varepsilon/2 - \varepsilon^2/8 + O(\varepsilon^3)$. A Taylor expansion around x gives

$$\begin{aligned}
\Phi\left(\frac{x}{\sqrt{a}}\right) &= \Phi(x) + \Phi'(x)x(a^{-1/2}-1) + \frac{\Phi''(x)}{2}x^2(a^{-1/2}-1)^2 + \frac{\Phi^{(3)}(c)}{6}x^3(a^{-1/2}-1)^3 \\
&= \Phi(x) + x\Phi'(x)\left(\frac{\varepsilon}{2} - \frac{\varepsilon^2}{8} + O(\varepsilon^3)\right) + x^2\frac{\Phi''(x)}{2}\left(\frac{\varepsilon^2}{4} + \varepsilon^\ni\right) + x^3\Phi^{(3)}(c)O(\varepsilon^3)
\end{aligned}$$

where c belongs to $(x, x/\sqrt{a})$. Since $x\Phi'(x)$ and $x^2\Phi''(x)$ can each be written $P(x)e^{-\frac{x^2}{2}}$ with P a polynomial of degree lower than 4, they are bounded on $\mathbb{R}$. The same holds for $x^3\Phi^{(3)}(c)$ since $|x^3\Phi^{(3)}(c)| \leq \sup_{x\in\mathbb{R}} |x^3\Phi^{(3)}(x/\sqrt{a})| \leq 1.2a^{3/2} \leq \infty$. We can therefore rewrite

$$\Phi\left(\frac{x}{\sqrt{a}}\right) = \Phi(x) + x\Phi'(x)\left(\frac{\varepsilon}{2} - \frac{\varepsilon^2}{8}\right) + x^2\frac{\Phi''(x)}{2}\frac{\varepsilon^2}{4} + O(\varepsilon^3)$$

uniformly in x. The same arguments lead to

$$\Phi\left(\frac{x}{\sqrt{b}}\right) = \Phi(x) + x\Phi'(x)\left(\frac{-\varepsilon}{2} - \frac{\varepsilon^2}{8}\right) + x^2\frac{\Phi''(x)}{2}\frac{\varepsilon^2}{4} + O(\varepsilon^3)$$

Combining these two equations and using $x\Phi'(x) = \frac{xe^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ gives the results. ■

Proof of Prop. 18: Since, $\frac{n^2}{k\sigma^2}\sigma^2_{n,+k} = \left(1+\frac{k}{n}\right)^{-1}$ and $\frac{n^2}{k\sigma^2}\sigma^2_{n,-k} = \left(1-\frac{k}{n}\right)^{-1}$, the result is a direct consequence of Lemma 26 when replacing ε by $\frac{k}{n}$. ■

B.5.4 Proof of Theorem 15

Proposition 28 *With the notations and under the conditions of Theorem 15*

$$\begin{aligned} F_-\left(\frac{\sqrt{k}\sigma x}{n}\right) &= \Phi(x) - \frac{kx}{2\sqrt{2\pi}n}e^{-\frac{x^2}{2}} + \sum_{j=1}^{m-2}\frac{P_j(-D)}{k^{j/2}}\Phi(x) + o\left(\frac{k}{n}\right) \\ F_+\left(\frac{\sqrt{k}\sigma x}{n}\right) &= \Phi(x) + \frac{kx}{2\sqrt{2\pi}n}e^{-\frac{x^2}{2}} + \sum_{j=1}^{m-2}\frac{P_j(-D)}{k^{j/2}}\Phi(x) + o\left(\frac{k}{n}\right) \end{aligned}$$

Uniformly in x, where D is the differential operator.

Proof. The two developments are obtained in the same way, we focus on the first one. It follows from Lemma 25 that

$$\begin{aligned} A &= \int_{-\frac{\sigma\sqrt{k}}{4\beta_m^{1/m}}}^{-\frac{\sigma\sqrt{k}}{4\beta_m^{1/m}}} \left|\frac{f_-\left(\frac{nt}{\sqrt{k}\sigma}\right) - e^{-\frac{t^2}{2}}\left(1-\frac{kt^2}{2n}\right)\left(1+\sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}\right)}{t}\right| dt \\ &\leq \frac{\delta(k)}{k^{\frac{m-2}{2}}}\int_{-\infty}^{\infty}(|t|^m + |t|^{3(m-1)})e^{-\frac{t^2}{4}} + \frac{K'_- k^2}{n^2}\int_{-\infty}^{\infty}(|t|^2 + |t|^{3m-2)})e^{-\frac{t^2}{4}} \\ &= O\left(\frac{\delta(k)}{k^{\frac{m-2}{2}}}\right) + O\left(\frac{k^2}{n^2}\right) = o\left(\frac{k}{n}\right) \end{aligned} \tag{B.12}$$

Since Cramér's condition holds, $\sup_{|t|\geq\beta_m^{-1/m}}|f_X(t)| \leq c < 1$. It follows then that

$$\int_{\frac{\sigma\sqrt{k}}{4\beta_m^{1/m}}}^{\sigma k^{\frac{m}{2}}}\frac{\left|f_+\left(\frac{nt}{\sqrt{k}\sigma}\right)\right|}{t}dt \leq \int_{\frac{\sigma\sqrt{k}}{4\beta_m^{1/m}}}^{\sigma k^{\frac{m}{2}}}\frac{\left|f\left(\frac{t}{\sqrt{k}\sigma}\right)\right|^k}{t}dt \leq \int_{\frac{1}{4\beta_m^{1/m}}}^{k^{\frac{m-2}{2}}}\frac{\left|f\left(t\right)\right|^k}{t}dt \leq k^{\frac{m-2}{2}}c^k = o\left(\frac{k}{n}\right) \tag{B.13}$$

The same holds for $e^{-\frac{t^2}{2}}\left(1-\frac{kt^2}{2n}\right)\left(1+\sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}\right)$. Finally, combining Eq. (B.12) and Eq. (B.13) gives

$$\int_{-k^{m/2}}^{k^{m/2}}\left|\frac{f_-\left(\frac{\sqrt{nt}}{\sqrt{k}\sigma}\right) - e^{-\frac{t^2}{2}}\left(1-\frac{kt^2}{2n}\right)\left(1+\sum_{j=1}^{m-2}\frac{P_j(it)}{k^{j/2}}\right)}{t}\right| dt = o\left(\frac{k}{n}\right)$$

Remark that $k^{-m/2} \sim n^{-\frac{m\alpha}{2}} = o(n^{-\frac{m}{m+2}}) = o\left(\frac{k}{n}\right)$. Using Theorem 22 with $T = k^{-m/2}$, we obtain:

$$F_-\left(\frac{\sqrt{k}\sigma x}{n}\right) = \left(1+\frac{kX^2}{2n}\right)(-D)\Phi(x) + \sum_{j=1}^{m-2}\frac{\left(1+\frac{kX^2}{n}\right)P_j(-D)}{k^{j/2}}\Phi(x) + o\left(\frac{k}{n}\right)$$

The term $\left(1+\frac{kX^2}{2n}\right)(-D)\Phi(x)$ of the right-hand side gives $\Phi(x)-\frac{kx}{2\sqrt{2\pi n}}e^{-\frac{x^2}{2}}$ when doing the inverse Fourier transform. The result then follows from $\frac{kX^2}{n}\frac{P_j(-D)}{k^{j/2}}\Phi(x)=o\left(\frac{k}{n}\right)$ uniformly in x. Replacing $1+\frac{kt^2}{2n}$ with $1-\frac{kt^2}{2n}$ in the proof gives the second expansion.

■

Proof of Theorem 15: The result is a direct consequence from Prop. 27. ■

B.5.5 Proof of Theorem 16

Remark: Cramér's condition is essential to ensure that the Edgeworth expansion of f_+ is valid up to the order m. If it does not hold, then Lemma 19 and 23 are still valid but $\int_\omega^t \frac{|f(t)|^k}{t}$ does not decrease exponentially fast anymore. We are limited to T or order $k^{1/2}$ in Theorem 22 so that only expansions of order 1 are available. But order 1 is not enough if n grows too fast compared to k.

Proposition 29 *With the notations and under the conditions of Theorem 16*

$$\begin{aligned}
F_-\left(\frac{\sqrt{k}\sigma x}{n}\right) &= \Phi(x)-\frac{kx}{2\sqrt{2\pi n}}e^{-\frac{x^2}{2}}+\frac{P_1(-D)}{k^{1/2}}\Phi(x)+o\left(\frac{k}{n}\right)\\
F_+\left(\frac{\sqrt{k}\sigma x}{n}\right) &= \Phi(x)+\frac{kx}{2\sqrt{2\pi n}}e^{-\frac{x^2}{2}}+\frac{P_1(-D)}{k^{1/2}}\Phi(x)+o\left(\frac{k}{n}\right)
\end{aligned}$$

Uniformly in x.

Proof. As for Prop. 27, the result is an application of Theo. 22. It follows from Lemma 25 that

$$\begin{aligned}
A &= \int_{-\frac{\sigma\sqrt{k}}{4\beta_3^{1/3}}}^{-\frac{\sigma\sqrt{k}}{4\beta_3^{1/3}}}\left|\frac{f_-\left(\frac{nt}{\sqrt{k}\sigma}\right)-e^{-\frac{t^2}{2}}\left(1-\frac{kt^2}{2n}\right)\left(1+\frac{P_j(it)}{k^{1/2}}\right)}{t}\right|dt\\
&\le \frac{\delta(k)}{k^{1/2}}\int_{-\infty}^{\infty}(|t|^3+|t|^6)e^{-\frac{t^2}{4}}+K'_-\frac{k^2}{n^2}\int_{-\infty}^{\infty}(|t|^2+|t|^7)e^{-\frac{t^2}{4}}\\
&= o(k^{-1/2})+o\left(\frac{k}{n}\right) \qquad \text{(B.14)}
\end{aligned}$$

Remark that, since $\alpha>2/3$, $k^{-1/2}\sim n^{-\alpha/2}=o(n^{\alpha-1})=o\left(\frac{k}{n}\right)$. Since Cramér's condition does not hold, we resort to Lem. 21 from which it follows that

$$\int_{\frac{\sigma\sqrt{k}}{4\beta_3^{1/3}}}^{\sigma\sqrt{k}\lambda(k)}\frac{\left|f_+(\frac{nt}{\sqrt{k}\sigma})\right|}{t}dt\le\int_{\frac{\sigma\sqrt{k}}{4\beta_3^{1/3}}}^{\sigma\sqrt{k}\lambda(k)}\frac{\left|f(\frac{nt}{\sqrt{k}\sigma})\right|^k}{t}dt\le\int_{\frac{1}{4\beta_3^{1/3}}}^{\lambda(k)}\frac{\left|f(t)\right|}{t}dt=o(k^{-1/2})=o\left(\frac{k}{n}\right) \qquad \text{(B.15)}$$

And the same holds for $e^{-\frac{t^2}{2}}\left(1-\frac{kt^2}{2n}\right)\left(1+\frac{P_1(it)}{k^{1/2}}\right)$. Combining Eq. (B.14) and (B.15) yields

$$\int_{-\sigma\sqrt{k}\lambda(k)}^{\sigma\sqrt{k}\lambda(k)}\left|\frac{f_-\left(\frac{nt}{\sqrt{k}\sigma}\right)-e^{-\frac{t^2}{2}}\left(1-\frac{kt^2}{2n}\right)\left(1+\frac{P_j(it)}{k^{1/2}}\right)}{t}\right|dt=o\left(\frac{k}{n}\right)$$

Since $\lambda(k)\to\infty$ as k, or equivalently n, goes to infinity, $\frac{1}{\sqrt{k}\lambda(k)}=o(k^{-1/2})=o\left(\frac{k}{n}\right)$. Theo. 22 then implies:

$$F_-\left(\frac{\sqrt{k}\sigma x}{n}\right)=\left(1+\frac{kX^2}{2n}\right)(-D)\Phi(x)+\frac{\left(1+\frac{kX^2}{n}\right)P_1(-D)}{k^{1/2}}\Phi(x)+o\left(\frac{k}{n}\right)$$

where D is the differential operator. The first term of the right-hand side gives $\Phi(x)-\frac{kx}{2\sqrt{2\pi n}}e^{-\frac{x^2}{2}}$. The result then follows from $\frac{kX^2}{n}\frac{P_j(-D)}{k^{j/2}}\Phi(x)=o\left(\frac{k}{n}\right)$ uniformly in x. Replacing $1+\frac{kt^2}{2n}$ with $1-\frac{kt^2}{2n}$ in the proof gives the second expansion. ■

Proof of Theorem 16: The result is a direct consequence from Prop. 28. ■

B.5.6 Proof of Theorem 17

Remark: As soon as X and Y have different expectations, $\Delta_{n,+k}$ is not centered anymore and the central limit theorem is enough to get the first order expansion of its distribution function. Up to a normalization constant, $\Delta_{n,+k}$ drifts away to $\pm\infty$, depending on the sign of $\mu_1-\mu_0$.

Following along the same lines as the proofs of Theorem 16, we first note that

$$f_+\left(\frac{nt}{\sqrt{k}\sigma_1}\right)\exp\left\{-it\frac{n}{n+k}\frac{\sqrt{k}(\mu_1-\mu_0)}{\sigma_1}\right\}=f_{Y-\mu_1}\left(\frac{n}{n+k}\frac{t}{\sqrt{k}\sigma_1}\right)^k f_{X-\mu_0}\left(\frac{\sqrt{k}t}{(n+k)\sigma_1}\right)^n$$

Using Lemma 24,

$$\left|f_{X-\mu_0}\left(\frac{\sqrt{k}t}{(n+k)\sigma_1}\right)^n-\left(1-\frac{\sigma_0^2kt^2}{2\sigma_1^2n}\right)\right|\leq K(t^2+t^4)\frac{k^2}{n^2}\quad\text{for}\quad|t|\leq\frac{\sigma_0\sqrt{n}}{4\beta_3^{1/3}}$$

And it comes from Lemma 19 that

$$\left|f_{Y-\mu_1}\left(\frac{n}{n+k}\frac{t}{\sqrt{k}\sigma_1}\right)^k-e^{-t^2/2}\left(1+\frac{kt^2}{n}\right)\left(1+\frac{\alpha_3}{6\sigma_1^6}\frac{(it)^3}{k^{1/2}}\right)\right|\leq\frac{\delta(k)}{k^{1/2}}(|t|^3+|t|^6)e^{-\frac{t^2}{4}}\quad\text{for}\quad|t|\leq\frac{\sigma_1\sqrt{k}}{4\beta_3^{1/3}}$$

Using the trick $|AB-ab|\leq|A(B-b)|+|b(A-a)|$ with $A=f_{X-\mu_0}\left(\frac{\sqrt{k}t}{(n+k)\sigma_1}\right)^n$, $B=f_{Y-\mu_1}\left(\frac{n}{n+k}\frac{t}{\sqrt{k}\sigma_1}\right)^k$, $a=1-\frac{\sigma_0^2kt^2}{2\sigma_1^2n}$ and $b=e^{-t^2/2}\left(1+\frac{kt^2}{n^2}\right)\left(1+\frac{\alpha_3}{6\sigma_1^6}\frac{(it)^3}{k^{1/2}}\right)$, it comes from $|A|\leq1$ and $|b|\leq K(1+|t|^5)e^{-\frac{t^2}{2}}$ that

$$\left|f_{X-\mu_0}\left(\frac{\sqrt{k}t}{(n+k)\sigma_1}\right)^n f_{Y-\mu_1}\left(\frac{n}{n+k}\frac{t}{\sqrt{k}\sigma_1}\right)^k-e^{-\frac{t^2}{2}}\left(1+\frac{kt^2}{2n}\left(2-\frac{\sigma_0^2}{\sigma_1^2}\right)+\frac{\alpha_3}{\sigma_1^3}\frac{(it)^3}{k^{1/2}}\right)\right|\leq$$
$$K\left\{\frac{k^2}{n^2}+\frac{\delta(k)}{k^{1/2}}\right\}(t^2+|t|^9)e^{-\frac{t^2}{4}}$$

for $|t| \leq B = \frac{\min(\sigma_0,\sigma_1)\sqrt{k}}{4\beta_3^{1/3}}$. It then follows that

$$\int_{-B}^{B} \left| \frac{f_{X-\mu_0}\left(\frac{\sqrt{k}t}{(n+k)\sigma_1}\right)^n f_{Y-\mu_1}\left(\frac{n}{n+k}\frac{t}{\sqrt{k}\sigma_1}\right)^k - e^{-\frac{t^2}{2}}\left(1 + \frac{k}{n}\frac{t^2}{2}\left(2 - \frac{\sigma_0^2}{\sigma_1^2}\right) + \frac{\alpha_3}{\sigma_1^3}\frac{(it)^3}{k^{1/2}}\right)}{t} \right| dt = o\left(\frac{k}{n}\right) + o(k^{-1/2}).$$

Lemma 21 combined to Theorem 22 then provides the following result:

$$\begin{aligned} F_+\left(\frac{\sqrt{k}\sigma_1 x}{n} + \frac{k}{n+k}\frac{\mu_1-\mu_0}{\sigma_1}\right) &= P\left\{\frac{n\Delta_{n,+k}}{\sqrt{k}\sigma_1} - \sqrt{k}\frac{n}{n+k}\frac{\mu_1-\mu_0}{\sigma_1} \leq x\right\} \\ &= \Phi(x) + \frac{k}{n}\left(2 - \frac{\sigma_0^2}{\sigma_1^2}\right)\frac{x e^{-\frac{x^2}{2}} e^{-\frac{x^2}{2}}}{2\sqrt{2\pi}} + \frac{\alpha_3}{6\sigma_1^3}\frac{(1-x^2)e^{-\frac{x^2}{2}}}{\sqrt{2k\pi}} \\ &\quad + o\left(\frac{k}{n}\right) + o(k^{-1/2}) \\ &= \Phi(x) + O(n^{-\beta}) \end{aligned}$$

uniformly in x, where $\beta = \min(\frac{\alpha}{2}, 1-\alpha)$. In addition, if x is bounded by some M, we further have

$$F_+\left(\frac{\sqrt{k}\sigma_1 x}{n}\right) = \Phi\left(x - \frac{\sqrt{k}(\mu_1-\mu_0)}{\sigma_1}\right)$$

which concludes the proof. ■

Chapter 4

Discussion and Prospects

4.1 Summary

Phylogenetic estimates are the result of an inference process. As such, even if the evolution model used in the inference was a perfect description of the real evolution process, the estimates would still be subject to a certain amount of variability. In layman's terms, even if the estimate is unbiased, its variance is still non negative. The phylogenetic estimate is a random variable (the randomness is inherited from the data at hand) and its fluctuations must be controlled as tightly as possible for the estimate to be accurate.

But the randomness of the observations in the sample can pass to the estimation process in several ways. The most obvious way is discussed in Sec. A and result from basic sampling theory. As the sample size n is not infinite, Cramer-Rao bounds and Fisher information theory (Kendall and Stuart, 1973) tell us that the estimates have a built-in variance, of order $n^{-1/2}$ in a parametric framework. However, the use of sampling theory requires the observations to be independent and identically distributed. Although, the independence assumption has been known for some time now to be a gross approximation (Bérard et al., 2008; Duret and Galtier, 2000), let us not question it yet. The sequential theory and several biological processes (recombination) strongly hints that the "identically distributed" is also far from obvious. Observations coming from different parts of the genome need not share a common evolutionary history and without this common history, it is hard to believe that the observations are identically distributed. Sampling theory and results presented in Chapter A are valid only for some portions of the chromosome in which observations are comparable. The procedure presented in Chapter B tests the distribution consistency of sequential data and can be used to assess whether observations coming from a series of observations are identically distributed.

4.1.1 Concentration Inequalities and Gaussian Approximation

As stated in Section 3.2 and Chapter A, for a given evolution model, topologies are compared to each other only via their likelihood scores. The aim of concentration

inequalities developed in Chapter A is to quantify, or rather upper-bound, the probability that the likelihood of a phylogeny wanders too far away from it expected value. Since phylogenies are ranked by their likelihood scores and chosen according to their rank, tight inequalities warrant an empirical ranking close to the true one. Since two very different phylogenies can have very similar likelihood scores, especially if the true topology is hard to resolve or encompass many short branches, this is not enough to guarantee that the phylogeny with highest (empirical) likelihood is the true one, or even has the same topology, but this is inherent to maximum-likelihood inference.

The two events studied in Chapter A are inversion events and concentration around the expected value. Concentration around the mean corresponds to the likelihood score being within margin α of its expected value whereas inversion events correspond to two topologies being in different orders when comparing their empirical likelihoods and true likelihoods. The probability of both this events depends on many quantities, including but not limited to the number s of species, the number n of observations, the true topology and evolution process (Q), α, etc. This should not come as a surprise. Indeed, s impacts the complexity of the model and the number of trees in the models, the true topology and evolution process determine the likelihood scores of all topologies and the resulting ranking and α is a confidence level. It is only natural for them to appear in the result.

As is usual for concentration inequalities, the results are non asymptotic and exact even for finite n. Also usual for concentration inequalities is the form of the upper bound: exponential decrease of the form e^{-Cn}. Although C should properly be written $C(\alpha, s, Q)$, it does not depend on n. Note that the exponential decrease is loose when compared to the rate obtained with the competing method of gaussian approximation (Sec. 2.4.3), which is of the form $e^{-C'n}$ with $C' > C > 0$. However, gaussian approximations are valid only under the strong assumption that the model is correctly specified. Considering the complexity of molecular evolution process, this is at best wishful thinking. Concentration inequalities, although looser than gaussian approximation, remain useful even when the true evolution process Q is not one of the usual evolution model, or even too complex to be described as a Markovian evolution model on a tree.

An intuitive way to understand this trade-off between tightness and versatility is in terms of worst case and average case. In the gaussian approximation, the probability of extreme observations is very low and all observations are close to an average observation. Because of this, the empirical likelihood score needs a significant fraction of observations to be extreme before significantly departing from its expected value. This happens only with extremely low probability (of order $e^{-C'n}$). Outside the gaussian approximation and for heavy-tailed distributions, extreme observations are still rare by definition but can have a much larger impact on the likelihood score than their gaussian counterpart. Therefore, the likelihood score needs a much smaller fraction of observations, and at the extreme only a few, to be extreme before significantly departing from its expected value. This happens with much higher probability (of order e^{-Cn} with $C < C'$) and is the main focus of Part II

4.1.2 The Pitfall of Increasing Sequence Length

Simplifications made in the test The stated goal of Chapter B is to study the distribution consistency of sequential data and yet the statistic we use, $\Delta_{n,+k}$, measures only modifications of the sample average. The choice of $\Delta_{n,+k}$ as a test statistic may seem crude at first sight. Indeed, a distribution can change in many ways while keeping a constant average. However, remember that in the maximum likelihood framework trees are scored by their likelihood score $\ell^T = \mathrm{E}_Q[Z] = \mathrm{E}_Q[\log \mathrm{P}(.;T,M)]$. We are therefore interested only in detecting changes in the distribution that affect this likelihood score, or equivalently that change the expected value of Z. Formally, a change from Q_0 to Q_1 is interesting if and only if $\mathrm{E}_{Q_0}[Z] \neq \mathrm{E}_{Q_1}[Z]$.

Here again, we need to distinguish two probability distributions. The first one is the probability distribution $P(.;T,M)$ induced by the evolutionary model (T,M). It is completely characterized by the tree T, the model M of sequence evolution and their associated parameters: branch lengths for the tree and parameter values for the model of sequence evolution. The second one is the sampling distribution Q, or the true probability distribution induced by the real evolutionary process. When both modeling and inference are accurate, we hope that Q is close to $P(.;T,M)$, in the sense that the Kullback-Leibler divergence $KL(P,Q)$ is small. However, there is no guarantee this is the case in practice (Goldman, 1993a,b). The crucial point here is that an observation X takes value x with probability $Q(x)$ and not $P(x;T,M)$.

The two distributions play very different roles. $P(.;T,M)$ is only used to compute the log-likelihood $\log P(x;T,M)$ of an observation x whereas Q is the sampling distribution, under which expectations are computed. Although Z has the very specific form $Z = \log P(X;T,M)$ and is thus not just any random variable, it is clear by noting $f(x) = \log P(x;T,M)$ and considering the random variable $Z = f(X)$, that changes in $\ell^T = E_Q[Z]$ occur only through changes in Q. Since $P(.;T,M)$ plays only a minor role when compared to Q, we neglected the specific form of Z and proposed a test to detect changes in the expectation $E_Q[Z]$ for any real-valued random variable.

Pros and cons of the non parametric test For most evolution models, $P(.;T,M)$ is a parametric distribution resulting from a Markov chain running on a tree (Sec. 2.2.2). It is both well behaved and can be completely characterized by a topology and a few parameters, typically of order $2s$ with s the number of species. Unfortunately, the same thing is not true of Q. We have little to no information about Q apart from its support, $\mathcal{A}^s = \{A,C,G,T\}^s$, and the only way to parametrize Q is as a multinomial distribution with (unknown) parameter $\theta = (\theta^x)_{x\in\mathcal{A}}$ of dimension $4^s - 1$. This parametrization of Q is so complex it seems difficult to use it to our advantage. Therefore, instead of restricting Q in the class of multinomial distributions to build a parametric test of consistency based on likelihood ratio, we opted for a non parametric test that works for any non-lattice distribution on $\mathbb{R}$. The drawback of this method is that the convergence speed is certainly not as fast as what could be achieved with a parametric test.

4.2 Further Work

Application to data set and other parameters Chapter B introduces a procedure to test the distribution of consistency of in sequential data. The test can be asymptotically calibrated to a given level. The calibration comes from asymptotic normality. Edgeworth expansions are used to correct first order departures from normality and guarantee that the probability of type I error is correct to the order $o(k/n)$. We would like to study how close this probability is to its designated target for several examples of distributions, including some significant departures from normality (asymmetry, high-skewness,etc). The next step would be to apply the procedure to examples for which the phylogenetic signal is well-known to be inconsistent along the sequence.

Finally, in order to simplify the problem, we reduced the trees to their likelihood scores as in section 2.4. Studying real-valued scores is more comfortable than studying topologies and allows precise evaluation of the range of expected shifts of the sample mean. We might however be interested in consistency of some other quantity: parameter of the evolutionary model, branch lengths, etc. It would be useful to check the validity of similar procedures applied to other quantities. In other words, can we test the consistency of κ along the sequence using the procedure described in chapter B ? How does the procedure behaves for more complex statistics than the distribution mean and can the test be calibrated in a similar way ?

Faster convergence rates A potentially promising way to achieve faster rates in the concentration inequalities is to condition on the absence of extreme sites. With the notations of Chapter A, the log-likelihood score of an observation is a random variable Z taking value $\log P(x;T,M)$ with probability $Q(x)$, where x is a value the observation can take. In the extreme case where the true evolution process is adequately described by a Markovian process on a binary tree, we can take $Q(.) = P(.;T,M)$. In such a case, if we list $0 \leq \theta^{x_1} := \theta_1 < \ldots < \theta^{x_N} := \theta_N \leq 1$ the possible values of $Q(x)$, Z takes value $\log(\theta_i)$ with probability θ_i. The closer θ_i is to 0, the higher the absolute value $|\log Q(\theta_i)|$ is. It is thus easily seen that the sites with biggest impact on the score (measured by $Q(\theta_i)$) are also those with lowest apparition probability (θ_i). Given the absence of extreme sites, or formally given $\{Z \geq \log Q(\theta_{i_0})\}$ for a carefully chosen i_0, the observations have a limited dispersion and are closer to their expected observations. This is in a sense a mean to get close to the gaussian approximation. The main problem here is the choice of i_0. For n observations, the probability of not observing a single extreme value is $(1-p*)^n$ where $p* = \sum_{i=1}^{i_0} \theta_i$ is the probability of observing *at least* one extreme value. For the conditioning to be interesting, i_0 needs to be finely tuned so that $p*$ decreases at least as $1/n$. Otherwise, $(1-p*)^n$ converges to 0 and we condition by a vanishing event. The remaining θ_i are then higher than K/n for some K and hopefully a faster rate of decrease can be achieved with this additional information. The intuitive rationale is that θ_{i_0} acts as a threshold for observations potentially extreme enough to affect the likelihood score and impede the correct ranking. That is why conditioning on their absence may yield faster rates. Of course, as the number of observations increases, some observations are not deemed extreme anymore, which is why θ_{i_0} decreases with n.

Another interesting development, related to conditioning would be to use dif-

ferent concentration techniques to upper bound the log-proba $\log \mathrm{P}(|\theta_i - \theta_{i,n}| \geq \varepsilon)$. For example, different inequalities can be used depending whether θ_i is small or large. Bennett's inequalities are well suited to small θ_i, which correspond to small variances whereas Hoeffding's inequalities are easier to manipulate for larger θ_i, which correspond to larger variances. Customized asymptotic developments of the rate function $h_p(\varepsilon) = (1-p-\varepsilon)\log\frac{1-p-\varepsilon}{1-p} + (p+\varepsilon)\log\frac{p+\varepsilon}{p}$ can also be used in place of Hoeffding's and Bennett's inequalities, following in the steps of Chapter A. Finally, the last development is a mixture of the two previous developments. It consists in neglecting the dependence of the (θ_n^x) and consider them simply as independent random variables. Noting that $\theta_n^x = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i=x\}}$, we know that $n\theta_n^x$ follows a binomial distribution $\mathcal{B}(n,\theta^x)$. Furthermore, all θ_n^x are correlated via the following correlation scheme: $\mathrm{Cor}(\theta_n^x,\theta_n^y) = -\dfrac{\theta^x\theta^y}{n}$. The approximation "$\theta_n^x$ and θ_n^y independent" is often used when estimating the number of classes in a population (Chao, 1984; Mao and Lindsay, 2007). Since $\dfrac{\mathrm{Cor}(\theta_n^x,\theta_n^y)}{\sqrt{\mathrm{Var}(\theta_n^x)\mathrm{Var}(\theta_n^y)}} = \sqrt{\dfrac{\theta^x\theta^y}{(1-\theta^x)(1-\theta^y)}}$, for small θ^x and θ^y the correlation is neglectable with regards to the variance of the estimators. Since θ^x and θ^y are typically very small in phylogenetic problems, the independence approximation is not outrageous. Under this approximation, each θ_n^x can be easily approximated by either a gaussian variable $\mathcal{N}\left(\theta^x, \frac{\theta^x(1-\theta^x)}{n}\right)$ if θ^x is larger than some K/n or a Poisson variable if θ^x is or order $1/n$. Under these approximations, the empirical likelihood score becomes a weighted sum of independent (gaussian or Poisson) random variables, whose distribution is fairly easy to characterize and control.

Including additional information Chapter A assumes that the analytical model is wrong as the true evolutionary process is too complex to be simply modeled by a continuous-time process running on a tree. The true distribution Q of the observations is thus parametrized in its most general form as a multinomial distribution with no structure in the probability vector. It is more realistic than the other way round but also less powerful. Indeed, for such Q and $P(.;T,M)$, the distance $|\ell_n^T - \ell^T| = |E_Q[Z] - E_{Q_n}[Z]|$ can only be bounded only through the quantities $\theta^x - \theta_n^x$, as we do in chapter A. The drawback of doing this is that all outcomes are dealt with in the same way; an observation x with apparition probability θ^x induces the uncertainty $(\theta^x - \theta_n^x)\log P(x;T,M)$ bounded by $|\theta^x - \theta_n^x| \times \|\log P(.;T,M)\|_\infty$ no matter whether θ^x is very small or close to 1/2. This quantity has variance $\theta^x(1-\theta^x)\log^2 P(x;T,M)$ bounded by $\theta^x(1-\theta^x)\|\log P(.;T,M)\|_\infty^2$ with the $\|\log P(.;T,M)\|_\infty$ term not depending on x. However, given that the most unlikely observations have very small probabilities $P(x;T,M)$, the difference between $\log^2 P(x;T,M)$ and $\|\log P(.;T,M)\|_\infty^2$ can be significant, especially for observations x with large $P(x;T,M)$. This bounding is thus suboptimal.

The problem here lies in the relationships between Q and $P(.;T,M)$. Without any further assumption, the distance and thus similarity between Q and $P(.;T,M)$ is vague at most. However, phylogenetics is mostly an inference problem. And remember from Chapter 2 that maximum-likelihood inference can be thought of as finding the model (T,M) which minimizes the Kullback-Leibler divergence $KL(Q,P(.;T,M))$. Under adequate inference, it is reasonable that Q is close to $P(.;T,M)$, or in other words $\theta^x = Q(x) \simeq P(x;T,M)$. In this case, the quantity $(\theta^x - \theta_n^x)\log P(x;T,M)$ has

variance roughly $(1-\theta^x)\theta^x \log^2 \theta^x$. The gain with respect to the previous bounding is of order $\log^2 \theta^x - \min_y \log^2 \theta^y$, and is, as expected, larger for x with large θ^x. External information about the similarity of Q and $P(.;T,M)$, such as an upper bound of the Kullback-Leibler or total variation distance between Q and $P(.;T,M)$ could thus be used to tighten the upper-bounds used to control $\ell^T - \ell_n^T$. Other refinements can also be proposed.

Assuming a correct analytical model A special case of additional information arises when the analytical model is correct, that is to say, there is a (T_0, M_0) such that $P(.;T_0,M_0) = Q$. Assume furthermore the model is identifiable, that is for all (T,M) and (T',M'):

$$P(.;T,M) = P(.;T',M') \Longrightarrow (T,M) = (T,M)$$

and that (T_0, M_0) is well separated from all other model:

$$\inf_{T \neq T_0} \inf_{M} KL\Big(P(.;T,M), P(.;T_0,M_0)\Big) > 0. \tag{4.1}$$

For a given n-sample corresponding to empirical distribution Q_n, the maximum likelihood estimates $(\hat{T}, \hat{M})$ of (T,M) satisfies:

$$\begin{aligned}(\hat{T},\hat{M}) &= \arg\max_{(T,M)} E_{Q_n}\left[\log P(X;T,M)\right] = \arg\max_{(T,M)} \ell_n^T \\ &= \arg\max_{(T,M)} \left(\ell_n^T - \ell^T\right) - KL(P(.;T,M), P(.;T_0,M_0))\end{aligned}$$

Under mild conditions on (T,M), such as:

$$\inf_{(T,M)} \min_x P(x;T,M) > 0,$$

it can be shown that $\ell_n^T - \ell^T \xrightarrow[n\to\infty]{} 0$ uniformly in (T,M) almost surely. It then follows from the separability condition (4.1) that the ML estimates are asymptotically consistent. The condition on (T,M) is satisfied as soon as transitions from any nucleotide to any other nucleotide happen with probability bounded away from 0 on any branch of any tree. In terms of evolutionary parameters, this translates to branch lengths of the tree bounded away from 0 and rate parameters also bounded away from 0. This condition is very mild because already necessary to ensure identifiability. It can also be used to make concentration inequalities tighter, as suggested in section 4.2 or to bound the probability of inferring the wrong topology (Steel and Szekely, 1999, 2002, 2006a).

In addition to being consistent, the ML estimates $(\hat{T}, \hat{M})$ have a simple interpretation in terms of Kullback-Leibler divergence; they minimize the divergence between Q_n and $P(.;T,M)$.

$$(\hat{T},\hat{M}) = \arg\max_{(T,M)} E_{Q_n}\left[\log P(X;T,M)\right] - E_{Q_n}[\log Q_n(X)] = \arg\min_{(T,M)} \left\{KL(P(.;T,M), Q_n)\right\}$$

The difference between the likelihood scores of $(\hat{T}, \hat{M})$ and (T_0, M_0) is given by:

$$E_{Q_n}[\log P(X;\hat{T},\hat{M})] - E_{Q_n}[\log P(X;T_0,M_0)] = KL(Q_n,Q) - KL(Q_n, P(.;\hat{T},\hat{M})) > 0 \tag{4.2}$$

Thanks to separability condition (4.1), if $KL(Q_n, Q)$ is small enough, the only way for $(\hat{T}, \hat{M})$ to be closer to Q_n than (T_0, M_0) is that $\hat{T} = T_0$. If the focus is on the topology rather than other estimates (branch lengths, transition rates), we can upper bound the probability of inferring a wrong topology simply by upper bounding $KL(Q_n, Q)$. We thus forget about trees and models for a time and focus only on the empirical distribution. The Kullback-Leibler distance between the empirical distribution $Q_n = (\theta_n^x)_x$ and the distribution $Q = (\theta^x)_x$ induced by (M_0, T_0) is:

$$\begin{aligned} KL(Q_n, Q) &= \sum_{x \in \mathcal{A}} \theta_n^x \log \frac{\theta_n^x}{\theta_x} \\ &= \sum_{x \in \mathcal{A}} \theta_n^x \log \left(1 + \frac{\theta_n^x - \theta^x}{\theta^x}\right) \\ &\leq \sum_{x \in \mathcal{A}} \frac{\theta_n^x (\theta_n^x - \theta^x)}{\theta^x}. \end{aligned}$$

Although $KL(Q_n, Q)$ is a complex random variable, it easily bounded by a much simpler random variable, which lead to further bounds. For example, since $n\theta_n^x$ follows the binomial distribution $\mathcal{B}(n, \theta^x)$, $E_Q[\theta_n^x(\theta_n^x - \theta^x)] = \mathrm{Var}(\theta_n^x) = \frac{\theta^x(1-\theta^x)}{n}$ from which it easily results:

$$E_Q\left[KL(Q_n, Q)\right] \leq \sum_{x \in \mathcal{A}} \frac{(1 - \theta^x)}{n} = \frac{\mathrm{Card}(\mathcal{A}) - 1}{n}.$$

Using Markov's inequality, we can bound $KL(Q_n, Q)$ and inject this bound in Eq. (4.2). We can also use the general inequality $KL(P, Q) > d_{TV}^2(P, Q)$, where d_{TV} is the total variation distance, to bound the distance $d_{TV}(Q_n, P(.; \hat{T}, \hat{M}))$ and inject it in the concentration inequalities to make them tighter, as suggested in section 4.2.

Finally we may shift the focus. All the work in chapter A dealt with bounding $\ell_n^T - \ell^T$ for a given tree T to appreciate how sensitive to sample size the likelihood score of that tree was. When the analytical model is correct, we can do the same thing with the correct tree T_0, *without* knowing it. Indeed, noting ℓ_n the likelihood score of the empirical distribution Q_n defined as:

$$\ell_n = E_{Q_n}[\log Q_n(X)] = \sum_x \theta_n^x \log \theta_n^x.$$

We remark that ℓ_n does not depend on a tree and is readily calculable from Q_n. Furthermore, the difference $\ell_n - \ell^{T_0}$ satisfies:

$$\ell_n - \ell^{T_0} = \sum_{x \in \mathcal{A}} \theta_n^x \log \theta_n^x - \sum_{x \in \mathcal{A}} \theta^x \log \theta^x = KL(Q_n, Q) + \sum_{x \in \mathcal{A}} (\theta_n^x - \theta^x) \log \theta^x \tag{4.3}$$

The term $\sum_{x \in \mathcal{A}} (\theta_n^x - \theta^x) \log \theta^x$ of Eq. (4.3) is exactly the one studied in chapter A when we specify the correct model, that is $(T, M) = (T_0, M_0)$. The term $KL(Q_n, Q)$ captures the additional randomness induced by the distribution under which likelihoods are computed, namely Q_n being inferred rather than specified beforehand. The two parts could be bounded to control the typical range of $|\ell_n - \ell^{T_0}|$ with little knowledge of T_0.

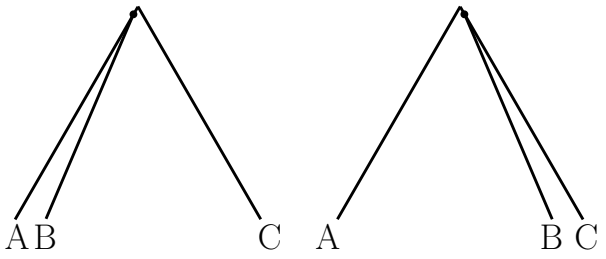

Figure 4.1: Left: Generating tree T_0 corresponding to the topology ((AB)C) with internal branch of length 0.01 and terminal branch of length 0.40. Right: Alternative tree T_1 corresponding to the topology (A(BC)) with internal branch of length 0.01 and terminal branch of length 0.40.

Dealing with multifurcating trees Up to now, we considered only model of the form (T, M) where T corresponds to a fully resolved topology, that is, a binary tree. However, it may happen that some internal branches of the tree are hard to resolve so that candidates corresponding to different resolutions of these branches, and thus to different topologies, are in a nutshell in terms of likelihood. The easiest and most visual example is perhaps that of short branches. Consider the trees represented in Figure 4.1. Under a simple Jukes-Cantor model for binary character, the likelihood scores of T_0, the correct tree, is $\ell^{T_0} = -1.3858$. The scores of ℓ^{T_0} and ℓ^{T_1} differ only by $\ell^{T_1} - \ell^{T_0} = 7.77 \times 10^{-5}$, four orders of magnitudes smaller that ℓ^{T_0}, and the variance of the likelihood of an observation is $\mathrm{Var}\left(\log P(X; T_0, M_0)\right) = \sigma^2 = 0.112$, a thousand time larger than the difference $\ell^{T_1} - \ell^{T_0}$. As a rule of the thumb derived from gaussian framework, n should be larger than $\frac{4\sigma^2}{(\ell^{T_0} - \ell^{T_1})^2} \simeq 8 \times 10^6$ to separate both T_0 from T_1, based on their likelihood score, with high confidence. And the problem can get worse when considering more complex trees with deep short branches.

In fact, it is well known that deep branches are intrinsically hard to resolve (Gronau et al., 2008; Mossel and Steel, 2005). Although the link between the length ε of the shortest branch in a tree and the number n of observations required by ML to resolve it correctly with high probability is still an open problem, the following relation was conjectured by Mike Steel during the PLGW01 in Cambridge in september 2007 Workshop (Phylogenetic Workshop 01 - Current Challenges and Problems in Phylogenetics):

$$n \propto c\frac{\log s}{\varepsilon^2}.$$

Since short branches require a large number of observations to be correctly resolved, the traditional way of doing phylogenetics by inferring fully resolved tree may be doomed for certain set of species which accumulate both short inner branches and a limited number of observations. Unlike some others fields in which the number of observations is limited only by the time, will and effort needed to gather them, there is a physical limit to the number of observations one can gather in molecular phylogenetics. "Observations" indeed correspond to nucleotides present in the genome of a set of species. As such, the number of observations can not exceed the size of those genomes. And even this is an optimistic overestimation. In practice and for various reasons such as lack of informativeness of some nucleotides, alignment issues,

incongruences in the phylogenetic signals, lack of independence and many others, the number of observations that qualify for inference after exhausting the genome is usually orders of magnitude smaller than the genome size. And there is little hope to ever increase it.

In such an instance where confidently resolving short branches requires longer sequences or more observations than will ever be available, it would be wiser to reduce our ambitions and settle for a more limited goal. Among the many ways to settle for less ambitious goals, a natural one is to consider only branches whose length exceeds some threshold depending on the available number n of observations. Indeed, if short branches can not be resolved with high probability until we have so many sites, thinking the other way round it is intuitive, with only so many sites at hand, to focus on those branches which can be resolved confidently and ignore or discard the others. This is done by relaxing the binary assumption and allowing multifurcating trees.

Binary trees correspond to models with the exact same dimension: same number of inner branches and thus of model parameters. However, multifurcating trees have less parameters. For example, a tree T with all but one binary inner nodes and a single ternary inner node has exactly one branch less than a binary tree T' with the same leaves. Depending on the complexity of the Markov model of sequence evolution M, (T', M) has at least one more parameter than (T, M), the branch length, and up to many more in for example a covarion-model with non-homogeneous stationary distribution (Lartillot et al., 2007). If T can be obtained by shrinking a branch of T', (T, M) and (T', M) are nested models which can be compared with classical likelihood ratio test (LRT). However, they are not nested in general and we must account for the difference in the number of parameters without resorting to LRT. We saw in Section 4.2 that phylogenetic inference can be thought of as model selection, with the model being completely characterized by a tree T and a Markovian evolution model M. We can thus use standard model selection criterion (AIC, AICc, BIC, etc) to select the best model (T, M). Note that traditional ML phylogenetic inference select the best model (T, M) from the set $\mathcal{M}^{bin}_{1,M} = \{(T', M') : T' \text{ is binary and } \quad M' = M\}$ in which M' is fixed and T is binary. Several tools developed by the community (Modeltest, ProtTest Posada and Crandall (1998)) to select the best evolution model can also be thought also as selecting the best model (T, M) from the set $\mathcal{M}^{bin}_{2,T} = \{(T', M') : T' = T\}$ in which T is a fixed complete binary topology. It is only natural to extend the model selection problem to a bigger set of models such as $\mathcal{M}_{1,M} = \{(T', M') : M' = M\}$ in which the evolution model is fixed but the tree need not be binary, $\mathcal{M}_{2,T} = \{(T', M') : T' = T\}$ in which the topology is fixed but not necessarily binary or even $\mathcal{M} = \{(T', M')\}$ in which no conditions is imposed to M nor to T. Although selection in such large sets of model is undeniably problematic, it is a promising way. Once a well supported multifurcating tree has been found and there is no well supported refinement of this tree, the multifurcating tree might be more relevant than any better resolved tree, in particular binary trees.

This is of course a complex inference issue. Recall from Section 2.3 that the tree space of unrooted binary tree is so big hill-climbing heuristics are the only affordable method to searching the maximum likelihood tree. The space of unrooted multifurcating trees with s leaves is of course larger than the space of binary trees with s leaves and the same problem arises. We must therefore devise moves to resolve or shrink

edges. A natural candidate would be to start from the binary tree and sequentially shrinking short edges until the penalized likelihood criteria selects a tree. But this is only one way to explore the space of multifurcating trees. Furthermore it always goes from the most to the less resolved trees, and other ways could be investigated.

Part II

Detecting Outliers

Summary

In this part, we use techniques coming from the robustness literature to study two potential sources of variability: outlier sites and rogue taxa. Outlier sites and rogue taxa are two of many sources that make the phylogenetic inference problem a difficult one. We describe in chapter 5 robustness analysis in phylogenetics, most notably bootstrap analysis, before focusing on outlier sites and rogue taxa.

Outlier sites impede the inference and belong to the larger class of influent sites whose inclusion/exclusion has high impact on the inferred tree. In chapter C, we propose a method to identify influent sites but independent lines of evidence remain necessary to distinguish between outliers and just influent sites. Using the phylogeny of fungi, we show that the strategy of discarding the most extreme outliers achieves more robust phylogenies.

Rogue taxa are the counterpart of outlier sites for taxa. In chapter D, we study an analogous method to identify influent taxa, with the same distinction between rogue and influent taxa. Using the phylogeny of placental mammal, we show that the method is successful in identifying well known rogue taxa.

We conclude part II in chapter 6, by discussing our results on robustness. We discuss refinements of the proposed methods and alternative interpretations of the results. We also discuss some limitations of our methods, potential ways to solve them and other techniques that could be used to study the robustness of phylogenetics trees.

Chapter 5

Introduction

5.1 The hardships of confidence study in phylogenetics

Most applications of phylogenetics to other fields of biology require accurate estimates. However, the inference process is complex and error-prone. The potential sources of error form a long list which most notoriously includes, but is by no means limited to, confounding processes – incomplete lineage sorting, alignment errors, inadequate taxon sampling, etc –, model misspecification, non-identifiability, phylogenetic incongruities, sampling errors, ...

Part I deals with sampling errors and phylogenetic incongruities along the sequence in a "well-behaved" framework. In this framework, errors occur either because of the sampling process or because different genes correspond to different trees thanks for example to incomplete lineage sorting, recombination or horizontal gene transfer. But apart from this, all "observations", that is to say sites along the sequence or equivalently columns of the alignment are well behaved in the sense that they are assumed to be independent and identically distributed. Independence is only an approximation and CpG sites, where the specific configuration of a cytosine-nucleotide C occuring next to a guanine-nucleotide G (p stands for the phosphate separating C and G) in the linear sequence of base induces mutation hotspots, provide a clear violation of the independence assumption. CpG-rich genomes will be interpreted, in a naive analysis, as rapidly evolving and lead to artificially long branch lengths and evolution rate estimates, or in the worst case to wrong topologies.

Worse, even if the characters did evolve independently, sequencing and aligning sequences is not an easy task (Sec. 2.2.1) and some sites are doomed to be erroneous and provide a fallacious signal. However, the inference method is blind to the genuine or fallacious nature of the signal and interprets both of them as genuine signal. Matsen (Matsen and Steel, 2007; Matsen et al., 2007) considered a phylogenetic mixture of two trees with the same topology but different branch lengths and proved that the mixture can mimic a tree with a different topology. Although the examples used in Matsen and Steel (2007) are a bit artificial and would be hard to resolve anyway, it strongly makes the point that conflicting signals in the alignment, even if they agree on the topology and differ only on the branch lengths, can be interpreted by the inference method as evidence for a tree with a different topology.

It is therefore essential, first, to quantify the uncertainty associated to the estimates and then to check whether the inference rests on the complete alignment or is mainly driven by a few sites. In the latter case, the estimate might not repeatable; spurious phylogenetic signal in a site is more likely to have a high impact on the estimates that in the former case, especially if the site is an inference-driving one. A simple way to understand this is perhaps in terms of probability estimation with noisy observations. Noting X the number of occurrences of an event A of probability p in a sample of size n, the ML estimator of p is simply $\hat{p} = X/n$. For moderate to large values of p, X is of order np and miscounting a few (m) occurrences of A has little impact on X and therefore on $\hat{p}$; the precision loss m/X is of order $1/n$. However, for small p of order K/n, X is of order K and miscounting a few occurrences of A has a high relative impact on X and therefore on $\hat{p}$; the precision loss m/X is of order m/K, much higher than $1/n$. In the latter case, the estimate is highly sensitive to each observation for which event A was counted.

The best way to assess whether the inferred tree is repeatable or not is to "repeat the experiment". It allows one to infer the distribution of the estimate when the model is too complex for explicit computation, and the actual estimate can then be evaluated as typical or atypical with regards to this distribution. However, in the presence of spurious signal, variability of the estimates is inflated and atypical estimates may be falsely deemed as typical. If spurious signal is caused by random errors and not by systematic bias, repeating the experiment changes both the spurious signal and the set of affected sites. Either way, if inference rests on the whole alignment, we expect spurious signals to be averaged out along the alignments and the estimate not to change drastically. However, if inference is dominated by a few sites only, changes in the spurious signal can not always be always be averaged out, especially when it affects inference-driving sites, and subsequently large changes of the estimates will occur more frequently than expected. Returning to the probability estimation problem; when the error on X takes value ± 2 with equal probability $1/2$, the estimate $\hat{p}$ changes by more than 50% over consecutive repeats with probability 0.07 (instead of 0.03 without error) for $(n, p) = (1000, 0.01)$ whereas it changes more than 50% over consecutive repeats with probability 0 for $(n, p) = (1000, 0.5)$. Changes less than 10% occur with probability 0.957 (instead of 0.958 without error) for $(n, p) = (1000, 0.5)$ and probability 0.15 (instead of 0.18 without error) for $(n, p) = (1000, 0.01)$.

It is however not always possible to "repeat the experiment". Evolution is a unique event involving extinct species and apart from very specific examples, such as fast evolving retroviruses (Bello et al., 2007), the hope of igniting or observing a new round of evolution in a reasonable time scale is illusory at best. The best substitute to repeats of the experiment is the use of independent data sets. Separate genes are ideal candidates. In an idealized vision of genomes evolution, different genes are independently subject to similar selection pressures and evolve according to the same generating process. They therefore represent exactly what we want: independent realizations of the same evolutionary process. But reality is of course a bit more complex and we must face several issues. First, separate genes need not evolve independently, as illustrated by duplicated genes (Duret, 2008) or interacting proteins (Pazos and Valencia, 2001; Sato et al., 2006). Gene duplication events correspond a single gene being duplicated and fixed in the population. Following duplication, the function of the replicates are initially redundant. If one duplicate undergoes a muta-

tion that knocks out its function, a mutation that would have been counterselected in the absence of a replicate may get fixed and release the mutant replicate from purifying selection while the other one is still subject to purifying selection to maintain its function (Eisen and Wu, 2002; Gu et al., 2005; Papp et al., 2003). Evolution of one copy thus depends on the other copy being functional or not and hence the two genes do not evolve independently. Interacting proteins are under coordinated evolution and have similar phylogenetic trees; hence leading to a reduction of the variability and overconfidence in the estimates. Second, independent genes need not correspond to the same gene trees. Consider two genes G_1 and G_2 and assume that G_1 underwent recent Horizontal Gene Transfer from distant species S_1 to species S_2 but not G_2. It means that the sequence of DNA was transferred from an individual of species S_1 to an individual of species S_2, *which is not its offspring*, and then fixed in species S_2 recently. The two versions of G_1 in S_1 and S_2 thus only diverged for a short time and present a higher level of similarity the two versions of G_2; species S_1 and S_2 appear closer in the gene tree of G_1 than in the gene tree of G_2, illustrating the disagreement between the gene trees. Finally, independently evolving genes with the same gene tree may still correspond to different evolutionary processes. Comparing a gene G_1 with the product of a gene duplication event G_2 may give different branch lengths; relaxation of selective pressures on the replicate is some but no all organisms allows for punctual bursts of evolution which translates to inflation of certain branches of the gene tree of G_2, but not of G_1.

The number of genes which can be used as independent replicates of the same "evolutionary experiment" is thus far more limited than the number of genes in the genomes and we must resort to other means, first to quantify the variability of the estimates and then to test the sensibility of the estimates to the particular data set used for inference.

5.2 Robustness Analysis

Bootstrap The standard way to quantify the uncertainty of the analysis in maximum likelihood phylogenetics is via bootstrap analysis. Bootstrap is simple and has a very intuitive rationale yet gives powerful results. Independent replicates of the data set used in the inference are very difficult or even nearly impossible to obtain. By contrast, resampling in the original data set to create fictional ones, called bootstrap data sets is fairly easy. We therefore replace independent replicates by bootstrap ones. The rationale is that, given there are enough observations in the original data set, the empirical distribution is a good approximation of the unknown true distribution, hence sampling either from the empirical distribution or from the true distribution should give very similar results. Of course, the sample drawn from the empirical distribution are not completely correct but they should display variations typical of a same size sample from the true distribution. We start by giving a general explanation of the bootstrap before considering how it is used in phylogenetics.

The bootstrap is a general-purpose tool analogous to jackknife and developed by Efron (1979). Its principle is best understood with Figure 5.1. Assume we have a n-sample $\{X_1, \ldots, X_n\}$ of i.i.d. observations drawn from distribution $F_0(\theta)$, which

depends on the unknown parameter θ. We compute the estimate $T(X_1, \ldots, X_n)$ of θ and would like to know the distribution of this estimate, or at least its variability. If F belongs to a known family of distribution and T is simple enough, we can compute the distribution of $T(\mathbf{X})$ and know how it depends on θ. For example, when F_0 is a normal distribution with mean θ and variance 1, and $T(\mathbf{X})$ is simply the sample mean (as in Fig. 5.1), we know precisely that $T(\mathbf{X})$ follows a normal distribution with mean θ and variance $1/n$; we have exact knowledge of the distribution of $T(\mathbf{X})$ for every value of θ.

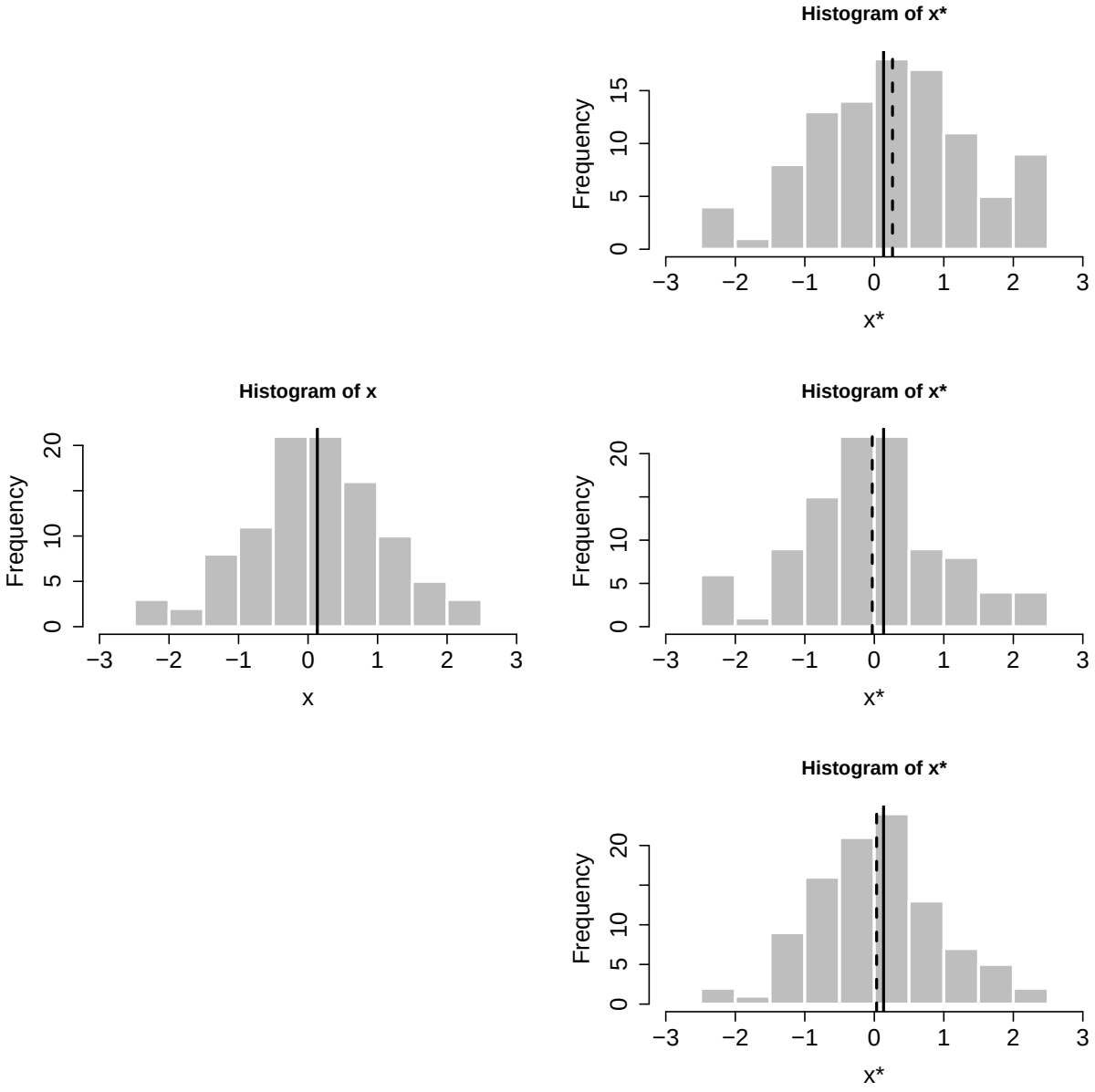

Figure 5.1: Left: The empirical distribution of independent observations is taken as an approximation of the unknown true distribution, in this case the standard normal distribution. Right: By drawing sample of size n (here $n = 100$) from it and analyzing them, we can approximate the variations that would be seen if we were able to draw new samples of that size from the true distribution. The parameter estimated is the mean of the distribution. The mean of the solid distribution is represented by a solid line, the mean of each bootstrap sample by a dotted line.

However, we may not know the distribution F_0 or $T(\mathbf{X})$ may be too complex for its distribution to be tractable and still want to quantify the typical variations of $T(\mathbf{X})$ if we repeated the analysis on an independent data set. Instead of using independent data sets, we use bootstrap data sets, that is, we draw new data sets not from the true distribution F_0 but from the empirical distribution F_1. To form a bootstrap data set, we draw new observations $\mathbf{X}^\star = X_1^\star, \ldots, X_n^\star$ uniformly from $X_1, \ldots, X_n$, independently and with replacement. If we sampled n observations without replacement, we would end up exactly with the original data set. Instead, drawing with replacement gives different data sets, in which some of the original observations are sampled once, twice, thrice and more or never. Estimating θ on the bootstrap data set $\mathbf{X}^\star$ gives a new estimate $T(\mathbf{X}^\star)$. The distribution of $T(\mathbf{X}^\star)$ can be estimated with arbitrary precision simply by drawing a large number B bootstrap data sets and computing $T(\mathbf{X}^\star)$ for each one. The variations of $T(\mathbf{X}^\star)$ should then be typical of the variations of $T(\mathbf{X})$. For many well-behaved distributions F and estimators T, theoretical results (Hall, 1984) assure that bootstrap is an accurate way to estimate the variability in the distribution of $T(\mathbf{X})$, given enough observations in the original data set and enough bootstrap data sets.

Bootstrap in phylogenetics To assess the uncertainty of phylogenetic estimates, bootstrap requires sequences of i.i.d. observations. Alignments are a matrix of species $\times$ character which can be read both ways. Species do not evolve independently but are instead related to each other through an unknown phylogeny. In fact, one of the major goal of phylogenetics is to discover this phylogeny. Characters are a much better candidate. Indeed, they are assumed to be independent and identically distributed in most, if not all likelihood based methods. Of course, the evolution of different characters may be related, as proved by CpG sites, so that the independence assumption is, strictly speaking, incorrect (Duret, 2006, 2008; Galtier and Duret, 2007). By reducing the effective number of observations but not the sample size, correlations reduce the variability of the bootstrap data sets. The estimates thus appear less variable than they really are (Felsenstein, 1985). Nevertheless, block-bootstrap (Künsch, 1989; Lahiri, 1999; Li et al., 2008; Otto et al., 1996) methods in which one draws blocks of sites instead of sites can cope with correlation between characters, hence bootstrapping sites is relevant.

In phylogenetics, bootstrap is mainly used to compute bootstrap values (*BP*) for branches. We assign the bootstrap value *BP* to a branch if it is present in a fraction *BP* of the bootstrap trees (Fig. 5.2). The *BP* value of a branch is intended to give an estimate of the amount of support the branch has. However, this value true meaning turns out to be shadowier than thought at first. Zharkikh and Li (1992) examined the statistical properties on unrooted 4-leaves trees and found that *BP* values underestimate the true support for large values of it. Hillis and Bull (1993) reached the same conclusion using simulation studies. They argued that *BP* as small as 70% may in fact indicate a significantly supported branch. Newton (1996) established a large deviation principle for the empirical distribution in a finite sample space, such as tree topologies, and show that the by-then well documented bias in *BP* values stems from dispersion effects in the joint distribution of sample and bootstrap sample. Felsenstein and Kishino (1993) agreed, in a reply to Hillis and Bull (1993), to the bias of *BP* values but argued that the problem does not lie in the bootstrap procedure itself but rather in the

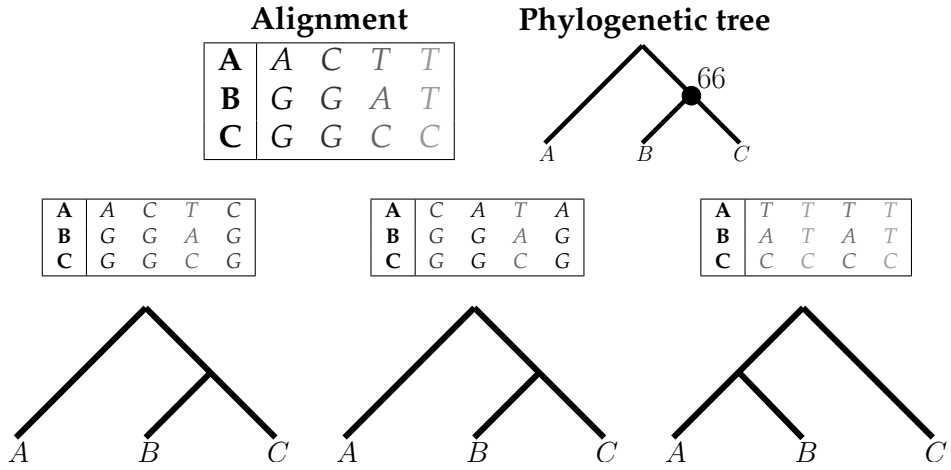

Figure 5.2: Example of bootstrap values. Top: Original data set and phylogeny estimate. Bottom: 3 bootstrap data sets and corresponding estimates. The node highlighted in the original estimate is present in 2 of the 3 bootstrap trees, hence its *BP* value is 2/3

use of *BP* as surrogates for the absence/presence of a clade. They interpret $1 - BP$ as a *p*-value for the test of the null hypothesis that the branch is not present. Efron et al. (1996) backed this interpretation and furthermore claimed that bootstrap is first-order correct. This means that the probability that the *BP* value is greater is than $1-\alpha$ should converge to α as the sequence length gets large, assuming the probability is calculated under the null hypothesis. However, Susko (2009) recently proved that *BP* values are not first-order correct, undermining the justification of bootstrap given by Efron et al. (1996). Notwithstanding any of these drawbacks, bootstrap remains a popular way to evaluate support for a clade or a branch in the maximum likelihood framework, as proved by its use in numerous studies.

5.3 Towards Specific Robustness Indicators

In our opinion, bootstrap is a simple yet powerful way to capture variability but has a poor resolution. As stated before, it is a simple non parametric approach to infer the variability of estimates when the model is too complex to allow for analytical calculation. In suitable conditions, it elegantly captures the variability of estimates. It is however unable to discover where it comes from or to break it down along the alignment. It will not tell us whether all sites contributes equally to the variability or whether a handful accounts for the lion's share. It is also useless to discover which regions of the alignment contribute most to the variability. Indeed, by the nature of bootstrap sampling, a character may or may not be sampled in a bootstrap data set. In fact, the probability it is not sampled is exactly $(1-1/n)^n$, or $1/e \simeq 0.3679$ with accuracy 1% as soon as $n \geq 100$, which is quite significant. Different bootstrap data sets are likely to be quite different; any observations has $e^{-2}\sum_{i=0}^{\infty}(i!)^{-2} = 30.8\%$ chance of being sampled the same number of times in different data sets; changes in the tree estimates are attributable to a large number of changes in the data set, which provides no way to assess which observations are essential and which ones are irrelevant. Bootstrap

aggregates all variations in the data set and is unable to distinguish between genuine and spurious variations.

The nature of bootstrap is to replace the sampling distribution with empirical distribution. However, the empirical distribution may be polluted by outliers. In the following, we adopt the following definition of outlier (Barnett and Lewis, 1994, p. 4): "An outlier is an entry in the data set that is anomalous with respect to the behavior seen in the majority of the other entries in the data set". Outliers may hinder the inference by pulling the estimates in one direction. By introducing additional variability in the data set, they may also artificially increase the variability of the estimate. Indeed, imagine we are interested in the mean of a distribution and have the following sample: $\{1, -1, 0, 1, 1, 0, -1, 0, 1, 10\}$. The last point (10) is clearly an outlier and has a huge influence on the empirical mean; its presence in the sample changes the estimate from 1.2 to 0.22 when excluded from the sample. 10 has probability $(1 - 1/10)^{10} = 0.348$ of being omitted from a bootstrap sample and 0.652 of being included at least once, which has a high impact on the mean of this bootstrap sample and thus on the variability of the bootstrap mean. The variability of the mean, estimated by bootstrap, shifts from 1.02 to 0.077 when removing 10 from the data set. Suppose an independent line of evidence allows us to characterize 10 as an aberrant point caused by measure errors. Cleaning it from the data set would yield more robust estimate.

Last but not least, bootstrap can also be seen as a way to check robustness of an estimate. Strictly speaking, a robust statistic is resistant to error in the results, produced by deviations from assumptions (Huber, 1981). *BP* values can be seen as a measure of robustness against distributional error. Indeed, if the sites are sampled from the empirical distribution F_1 instead of the true distribution F_0, the model assumption that observations are distributed according to F_0 is clearly violated (unless the original data set is so extraordinary that the empirical distribution coincides with the true distributions). *BP* value of a branch is then the resistance of this branch to changing the sampling distribution from F_0 from F_1. There are however at least another kind of desirable robustness: resistance to outliers. A statistic is outlier resistant if it is not affected by a very small fraction of outliers. Bootstrap treats all sites the same way so that outliers can not be specifically targeted. Furthermore, the difference between two bootstrap data sets is typically large — assuming all sites are different, each site has probability $(1 - 1/n)^n[1 - (1 - 1/n)^n] \simeq e^{-1}(1 - e^{-1}) = 0.232$, with accuracy 1% as soon as $n \geq 22$, to be sampled in one data set and not in another. Difference between a bootstrap data set and the original one is thus almost never limited to "a very small fraction of outliers" and bootstrap is not the proper way to test whether the inference method is outlier resistant.

Detecting Outliers Going back to our definition of outliers: "An outlier is an entry in the data set that is anomalous with respect to the behavior seen in the majority of the other entries in the data set", we first need to define the behavior seen in the majority of the other entries. We adopt a strategy based on our interest in the phylogenetic estimates and consider outliers, or influent sites, as sites which unduly affect the inference.

Influent sites can be divided into two categories: genuinely influent sites which

provide a strong signal to the inference and erroneous sites which just hinder the inference and introduce undeserved variability in the estimates. Even tough we must use independent lines of evidence to distinguish the former from the latter, removing erroneous sites from the data set should thus produce less variable, more robust trees. And the first step in this direction is of course to identify influent sites. The basic tools to detect such influent sites are influence functions and sensibility curve (Efron, 1979; Hampel, 1974b).

The influence function gives us an idea of how a estimator behaves when we change exactly one point in the sample. Formally, let A a convex subset of the set of all finite measures on $\mathbb{R}^d$ ($d \geq 1$) and $F \in A$ be a probability distribution. Suppose we want to estimate a parameter $\theta \in \Theta$ of $F \in A$ and let $T : A \to \Theta$ be an estimator such that $\forall \theta \in \Theta, T(F_\theta) = \theta$. To evaluate the importance of an additional observation $x \in \mathbb{R}^d$ we define:

$$IF_{T,F}(x) = \lim_{t \to 0^+} \frac{T(t\delta_x + (1-t)F) - T(F)}{t}$$

which measures the influence of an infinitesimal perturbation of the functional $T(F)$ along direction δ_x, the Dirac mass on x. If $IF_{T,F}$ is bounded uniformly in x, the estimate is resistant to outliers and extreme values, not matter how extreme.

Of course F is usually unknown so that we need an empirical version of the influence function which makes no such assumptions on the model. Furthermore, there are many phylogenetic estimates, some continuous (branch lengths, evolution model parameters,...) and some discrete (tree topology). We thus need to choose a relevant functional that incorporates many parameters. We present in Chapter C an adaptation of Influence Functions to phylogenetics, that allows one to detect influent sites. We then apply the method to a data set of fungus (158 taxa, 1026 sites) and display evidence that the two most influent sites are indeed outliers that hinder the inference by strongly pulling the tree in one direction. We show that "cleaning" the data set from as few as these two points result in more stable and better supported phylogenies.

Taxon Influence Outliers are not limited to sites. An alignment is a species $\times$ sites matrix and can be read both way. Jackknifing sites as we do in Chapter C to detect outliers has solid statistical foundations because the sites are i.i.d. Species, by constrast, are not and the whole point of phylogenetics is to discover how they are related to each other. With the usual models of sequence evolution, the only way for two species to be independent is to have diverged an infinitely long time ago; only then is the dependence induced by the Markov model of sequence evolution completely erased from the sequences. But this corresponds to a pathological tree with a branch of infinite length. Even if such a tree made some sense, alignment and tree inference are notoriously difficult for highly diverged sequences. This peculiar case is thus of little practical interest. The statistical framework for jackknifing species is far more complex than its equivalent for jackknifing sites.

However species sampling, or taxon sampling as will be used in Chapter D, has its importance in the inference. Swofford et al. (1996) argues that adequate taxon sampling for the problem of interest is one of the four primary factors for accurate phylogenetic estimates, on par with enough sequence data. The most striking

illustration is perhaps the decision to include or not an outgroup. The outgroup is a taxon or set of taxa distant enough from the rest of the taxa (the ingroup) to ensure than the most recent common ancestor (MRCA) of the ingroup is more recent that the MRCA of all the taxa, and therefore to root the phylogeny. It is known that the inclusion of the outgroup to the analysis may disrupt the ingroup phylogeny (Holland et al., 2003; Shavit et al., 2007). Another bias has been well documented since Felsenstein (1978); when two non-adjacent taxa share many character states along long branches because of convergence, some inference methods often interpret such similarity as homology. The resulting inferred tree displays the two taxa as sister taxa, attributing the shared changes to a branch joining them. This effect is termed Long Branch Attraction (LBA) and causes some methods, most notably parsimony, to be inconsistent, meaning they converge to an incorrect tree as the number of characters used in the data set increases. But extreme sensitivity of the estimates to taxon inclusion has no reason to be limited to outgroup or taxa at the end of long branches. Heath et al. (2008) even recommends that "analysis of sensitivity to taxon inclusion [...] should be a part of any careful and thorough phylogenetic analysis".

And there is still more. In addition to extreme sensitivity, phylogenetic analysis of few taxa (but each represented by many characters) can be subject to strong systematic biases, which in turn produce high measures of repeatability (such as bootstrap proportions) in support of incorrect or misleading phylogenetic results (Heath et al., 2008; Rokas and Carroll, 2005; Rokas et al., 2003). Jackknifing species is therefore useful to detect systematic bias, to which resampling methods are blind, but its theoretical properties when assessing the variability of the estimates are still to be uncovered.

We propose in Chapter D a procedure to describe and measure the influence of taxon sampling on the tree at the individual taxon level by jackknifing species. Influent taxa strongly affect the estimates when included or excluded for the analysis. Like outliers, influent species can be either influent for a reason and improve accuracy by reducing systematic bias or just be rogue taxa that hinder the inference by introducing bias. We then apply the procedure to a the placental mammal phylogeny (68 taxa, 3658 sites). The results provide evidence that rodents in general, and guinea pig in particular, are influent taxa, corroborating previous findings in the literature that they are rogue taxa. Furthermore, most branches of the inferred tree are highly resistant to taxon sampling, providing evidence that the bias induced by taxon sampling acts, in this example, only on a few branches.

Appendix C

Influence Function

This section is a modified version of the article *Influence function for robust phylogenetic reconstructions* published in Molecular Biology and Evolution.

A. Bar-Hen, M. Mariadassou, M.-A. Poursat and P. Vandenkoornhuyse, Influence function for robust phylogenetic reconstructions, *Molecular Biology and Evolution*, **25**:869-873, 2008

Abstract Based on the computation of the influence function, a tool to measure the impact of each piece of sampled data on the statistical inference of a parameter, we propose to analyze the support of the maximum likelihood tree for each site. We provide a new tool for filtering datasets (nucleotides, amino acids and others) in the context of maximum likelihood phylogenetic reconstructions. Because different sites support different phylogenic topologies in different ways, outlier sites, *i.e.* sites with a very negative influence value, are important: they can drastically change the topology resulting from the statistical inference. Therefore, these outlier sites must be clearly identified and their effects accounted for before drawing biological conclusions from the inferred tree.

A matrix containing 158 fungal terminals all belonging to Chytridiomycota, Zygomycota and Glomeromycota is analyzed. We show that removing the strongest outlier from the analysis strikingly modifies the maximum likelihood topology, with a loss of as many as 20% of the internal nodes. As a result, estimating the topology on the filtered dataset results in a topology with enhanced bootstrap support. From this analysis, the polyphyletic status of the fungal phyla Chytridiomycota and Zygomycota is reinforced suggesting the necessity of revisiting the systematics of these fungal groups. We show the ability of influence function to produce new evolution hypotheses.

Keywords Influence function, Phylogenetic, Maximum likelihood, Tree stability

C.1 Introduction

Phylogenetic methods are used in many diverse fields, including molecular evolution, virology and ecology. Maximum likelihood is one of the most popular. It is based on the adoption of an explicit DNA or protein sequence evolution model. Depending on the complexity of the model, the inferred tree can be very dependent on randomly ocurring peculiarities in the dataset, thus, its robustness must be assessed. The most commonly used test of reliability of an inferred tree is the bootstrap (Efron, 1979; Felsenstein, 1985), though the simulation output is, unfortunately, rarely examined to determine whether their conclusions are only driven by a few peculiar sites.

Empirical research in many areas of statistics gives high priority to detecting outliers. Indeed, outliers have a strong effect on the results of a statistical analysis and can even invalidate conclusions drawn from them. In molecular phylogenetics, every site takes part in the inference of a phylogenetic tree. But how stable is the inferred tree? In other words, are there any sites that drive the tree topology thus inducing change(s) when deleted? Does the support of a branch rest on an atypical segment of the DNA sequence? Drawing valid conclusions from a phylogenetic tree requires to control these outlier sites. Although the classical emphasis is to minimize the influence of such sites, the most interesting aspect might be to *detect* them. Influence functions, introduced by Hampel (1974a) as a measure of the impact that each piece of sampled data has on the statistical inference, are helpful to detect such influential segments of sequence. In this paper we make use of the influence

function concept to obtain influence diagnosis in phylogeny. Various other uses of the influence function can be found in Huber (2004), and the relationships between jackknife and influence function were proved in Miller (1974).

Resampling techniques are the most widely used approaches to assess the stability of inferred trees but there are other approaches that have been used to assess robustness in the context of phylogenetic analyses. For example, Archibald and Roger (2002) used a likelihood ratio test for scanning DNA sequence alignments to detect regions of incongruent phylogenetic signals, such as those influenced by recombination. Blouin et al. (2005) presented a simulation study in which they evaluated the robustness of evolutionary site-rate estimates for both small and phylogenetically unbalanced samples.

Since we want to characterize the influence of each site on the likelihood, it is crucial to study them one at a time. Let T be the tree that maximizes the likelihood of the whole dataset and $T^{(h)}$ be the tree that maximizes the likelihood of the jackknife sample obtained when removing site h from the original dataset. By comparing T to each $T^{(h)}$, we study the impact of each site on T and can relate the stability or lack of a stability of a clade to a particular site or set of sites. We also define the outlier sites as those whose influence values are the greatest. Outlier sites may arise from biological well-know characteristics which result in evolution schemes not taken into account by the evolution model, such as the nature of mutation of GC-content for a given nucleotide dataset. Taking a further step toward robustness, we order the sites in the original dataset from strongest outlier to weakest outlier and remove them one at a time, starting with the strongest outlier. Doing so, we obtain a sequence of samples, each one shorter than the previous one by exactly one nucleotide, from which the corresponding sequence of trees is inferred. Assuming that major causes of disruption and thus instability disappear along with the strongest outlier, we expect a stable tree to arise from this sequence. The main issue is then: how many outliers must be removed before the inferred tree becomes robust?

C.2 Methods

Definitions and notations Let us consider s homologous nucleotide sequences that consist of n nucleotide sites to construct a tree. Let $\mathbf{X} = (X_{pq})$ be the $s \times n$ matrix of data where X_{pq} is T, C, A or G and denotes the state of the qth site in species p. Let $\mathbf{X}_h = (X_{1h}, \ldots, X_{sh})'$ be the data at the hth site. The superscript $'$ denotes the transpose operator.

Assuming a substitution model and independently evolving sites, the log-likelihood of a given tree T is:

$$l_T(\boldsymbol{\theta_T}|\mathbf{X}) \;=\; \sum_{h=1}^{n} \log f_T(\mathbf{X}_h|\boldsymbol{\theta_T}) \tag{C.1}$$

where $f_T(\mathbf{X}_h|\boldsymbol{\theta_T})$ is the probability to observe pattern, *i.e.* alignment column, $\mathbf{X}_h$ at the homologous site h. We note that the log likelihood divided by the sample size, $l_T(\boldsymbol{\theta_T}|\mathbf{X})/n$, can be regarded as an unbiased estimator of the expected log likelihood

per site. Even if the sites are correlated, it is an unbiased estimator of the expected log-likelihood per site, under mild assumptions on the correlation structure (*e.g.* ergodicity of the Markov Chain modeling the correlation) (Bar-Hen and Kishino, 2000).

Given the topology describing the branching order, the log likelihood is expressed in terms of the transition probabilities computed from the evolution model. The vector $\boldsymbol{\theta}_T$ denotes the set of unknown parameters such as the branch lengths of tree T and the substitution rate of the evolution model. We refer to Bryant et al. (2005) for an up-to-date review on maximum likelihood techniques for phylogenetics.

Influence function for phylogeny. We adapt the concept of influence function to the context of phylogenetics. To a given alignment $\mathbf{X} = (\mathbf{X}_h)_{h=1,\dots,n}$, we associate the log-likelihood statistic:

$$S(F_n) = \frac{1}{n}\sum_{h=1}^{n} \log f_T(\mathbf{X}_h|\boldsymbol{\theta_T})$$

with $f_T(X|\boldsymbol{\theta_T})$ defined in equation C.1 and where T is the tree maximizing the likelihood of $\mathbf{X}$.

The effect of deleting site $\mathbf{X}_h$ can be measured by its influence value $IF_{S,F_n}(\mathbf{X}_h)$:

$$IF_{S,F_n}(\mathbf{X}_h) \quad = \quad (n-1)\Big(l_T(\boldsymbol{\theta_T}|\mathbf{X}) - l_{T^{(h)}}(\boldsymbol{\theta_{T^{(h)}}}|\mathbf{X}^{(h)})\Big) \tag{C.2}$$

with $\mathbf{X^{(h)}}$ representing all the sites of $\mathbf{X}$ but $\mathbf{X_h}$ and $T^{(h)}$ defined in the same way as T as the tree maximizing the likelihood of $\mathbf{X^{(h)}}$. The value $IF_{S,F_n}(\mathbf{X}_h)$ gives the (scaled) change in average likelihood resulting from removing site $\mathbf{X}_h$. If a site has a positive value, this means that the parameters estimated on all sites, including the new one, has a higher likelihood than the parameters estimated on all sites but the new one. And the opposite if a site has a negative value.

The most interesting property of equation C.2 is the possibility to characterize the sites with a strong influence, *i.e.* sites for which $IF_{S,F_n}(\mathbf{X}_h)$ is either very positive or very negative. A very positive influence value implies that the site strengthens the support for topology T whereas, a very negative value implies that the site weakens the support of topology T. In real case dataset, and under our assumption that only a few sites disrupt the robustness of the inferred topology, we expect to find many sites with small positive influence value and a few sites with large negative influence value. Therefore, we focus on sites with very negative influence value and call them *outlier sites*.

Stability of the maximum likelihood tree among trees maximizing the likelihood of pseudo-samples. The bootstrap is the most popular method in phylogenetics to assess the uncertainty of the inferred tree. Using pseudo-samples, P-values are computed for the branches of the tree. These P-value are intended to estimate the support provided by the data to a clade. They can be used to build a majority-rule consensus tree in which only clades with a P-value greater than 0.5 appear. The jackknife and influence function provide additional information to the stability of clades. Mainly, they relate the stability of a clade to certain particular sites. Thus,

original information can be extracted. For example, do the outliers have a specific nucleotide content?

Bootstrap analysis, just like any statistical analysis, is sensitive to individual observations. In a phylogeny analysis, questions such as "would the support of that clade differ if these sites were discarded from the analysis?" or "are the clades sensitive to the considered sample?" often arise. To answer them, it is important to focus on the effect of individual sites on bootstrap values. Empirical influence values are useful in this context, as they can identify nfluential sites (*i.e. outliers*.

Let $X_1, \ldots, X_n$ be random variables with common distribution function (df) F on $\mathbb{R}^d$ ($d \geq 1$). To simplify notations, we use distribution function and probability measure indifferently: F is either one or the other. Suppose that we are interested in a parameter that can be expressed, as often in statistics, as a functional $S(F)$ of the generating df, S being defined on the space $\mathcal{F}$ of df's.

To evaluate the importance of an additional observation $x \in \mathbb{R}^d$, we can define, under conditions of existence, the quantity

$$IF_{S,F}(x) = \lim_{\epsilon \to 0} \frac{S\Big((1-\epsilon)F + \epsilon\, \delta_x\Big) - S\Big(F\Big)}{\epsilon} \tag{C.3}$$

which measures the influence of an infinitesimal perturbation on the functional $S(F)$ along the direction δ_x (Efron, 1979). δ_x is the Dirac measure which concentrates the whole probability mass 1 on the point x. The influence function $IF_{S,F}(x)$ is defined pointwise by C.3, if the limit exists for every x.

Usually F is unknown, so that one has to estimate it by the empirical distribution function defined from the sample as:

$$F_n = \frac{1}{n} \sum_{h=1}^{n} \delta_{X_h}.$$

The natural estimator of $S(F)$ is then $S(F_n)$ and the empirical version of the influence function is obtained from C.3 by replacing F with F_n. The particular values $IF_{S,F_n}(X_h)$ are called the empirical influence values.

There is a strong connection between the influence function and the jackknife (Efron, 1979; Miller, 1974), which is a statistical technique for empirically estimating the variability of an estimator. The jackknife involves dropping one observation from the sample at a time and calculating the corresponding estimate each time. Let $F^{(h)}_{n-1} = \frac{1}{n-1} \sum_{j, j \neq h} \delta_{X_j}$ be the empirical df calculated with X_h omitted from the data. Then, $F_n = \frac{n-1}{n} F^{(h)}_{n-1} + \frac{1}{n} \delta_{X_h}$ and a numerical approximation of $IF_{S,F_n}(X_h)$ can be obtained using $\epsilon = -\frac{1}{n-1}$:

$$\begin{aligned} IF_{S,F_n}(X_h) &\approx \frac{S\Big((1-\epsilon)F_n + \epsilon\, \delta_{X_h}\Big) - S\Big(F_n\Big)}{\epsilon} \\ &= (n-1)(S(F_n) - S(F^{(h)}_{n-1})) \\ &= S^*_{n,h} - S(F_n) \end{aligned}$$

where $S^*_{n,h} = nS(F_n) - (n-1)S(F^{(h)}_{n-1})$ are the pseudo-values of the jackknife, *i.e.* the estimated values of $S(F)$ computed on $n-1$ observations (Miller, 1974).

An alternative to influence function to measure the impact of site $\mathbf{X}_h$ on the inference of a statistic S is *jackknife-after-bootstrap*: the value of S over the whole sample is compared to the values $S^*_1, \ldots, S^*_B$ obtained from bootstrap samples where $\mathbf{X}_h$ does not occur. However, the computional time involved in most maximum likelihood techniques makes it demanding, in time and in computer resources, to perform bootstrap analyses. In addition, influence functions are anchored in a more classical framework. Therefore, we favored influence function over jackknife-after-bootstrap.

C.3 The dataset

The influence function of each site was computed from an alignment of the SSU rRNA gene (1026 nucleotides) over 157 terminals (*i.e.* 157 rows), all fungi belonging to the phyla Chytridiomycota, Zygomycota, Glomeromycota plus one outgroup to root the tree, *Corallochytrium limacisporum*, a putative choanoflagellate. This alignment, previously published in Vandenkoornhuyse et al. (2002) was chosen to satisfy different criteria: (*i*) enough variation accumulated to clearly resolve the phylogenetic topology (*ii*) a very low number of detectable homoplasic events (*iii*) a strong monophyletic group (*i.e.* Glomeromycota) (*iv*) a highly polyphyletic group (*i.e.* Zygomycota) (*v*) one group with uncertainties about phylogenetic affinities (*i.e.* Chytridiomycota).

C.4 Results and Discussion

In this paper, we focused on the detection of influential sites (*i.e.* outliers) for the maximum likelihood tree of fungi belonging to the phyla Chytridiomycota, Zygomycota, Glomeromycota. The idea developed here is that computing influence values helps to detect outliers for the proposed model of evolution and to compute a more robust tree.

The influence function of each site was computed from an alignment containing 157 fungal terminals and 1026 nucleotide sites (*i.e.* 1026 columns and 157 rows) (see dataset section).

We first performed a maximum likelihood estimation of the phylogeny of the 158 sequences using the PHYML program (Guindon and Gascuel, 2003). The maximum likelihood (ML) tree T was constructed with the General Time Reversible (GTR) model (Felsenstein, 2004). Furthermore, we have evaluated the fit to our data of different models of nucleotide substitution (including HKY, F81, JC, ...) using "modeltest" (Posada and Crandall, 1998) `http://darwin.uvigo.es/software/modeltest.html` and confirmed the validity of the choice of the GTR model. The tree presented in supporting online material is in accordance with previously published trees (Van-

denkoornhuyse et al., 2002) and provides a result congruent to the maximum parsimony tree.

We used a R script, (avalaible upon request to the corresponding author) to compute the influence values C.2 for each of the 1026 sites of the alignment. Each influence value is computed by removing one site h from the whole dataset, computing the ML tree $T^{(h)}$ on the obtained jackknife-sample, and taking the difference between the mean likelihood of a site under the ML tree T and under $T^{(h)}$. We found out that certain sites have very negative influence values, *i.e.* that removing them strongly worsens the likelihood of the ML tree (Fig. C.1). Furthermore, some the $T^{(h)}$ were quite different from T. In other words, when removed, some sites significantly modified the inferred tree. Fig. C.1 plotted, for each site h, the number of internal nodes of the ML tree T not found in tree $T^{(h)}$. This proves that a change in the likelihood of a sample reflects a change in the underlying ML topology: change of topology and change of likelihood are strongly connected.

When removing a site, between 11 and 32 internal nodes of the ML tree were affected. Fig. C.1 showed an average of 15 nodes affected by removing only one site. These nodes were related to terminals with high homology within unresolved clades, *i.e.* not well supported by the ML tree. Some areas contained the strongest outliers which were not uniformly distributed along the sequence.

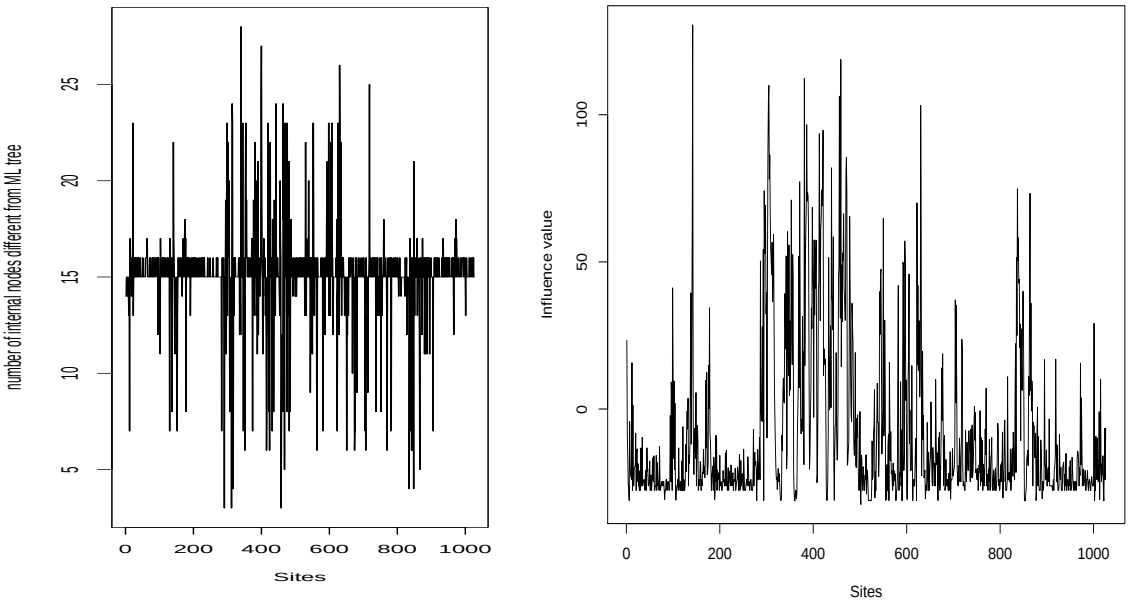

Figure C.1: (A) number of internal nodes different from the ML-GTR guide tree (all data included) when removing one site only from the dataset. and (B)influence values when removing each of the single sites (*i.e.* one column only) from the dataset (1026 columns in total)

For example, the most infuential site (*i.e.* strongest outlier) (position 142 on the dataset) corresponded to a highly variable site. To visualize the position of this particular site we computed the most probable RNA secondary structure (RNA folding) using a method based on thermodynamic principles (Zucker et al., 1999) (mfold at `http://www.bioinfo.rpi.edu/applications/mfold/`). From 2 different sequences selected randomly, and using different temperatures and different salinities we al-

ways found that the strongest outlier is on a small loop (5 nucleotides) carried by a conserved hairpin (figure not shown, available on request).

In order to achieve a more robust tree, we removed the strongest outliers from the analysis. If the outliers indeed disrupt the inferred topology, we expect that, after discarding enough of them, the inferred tree will not be over sensitive to the sample anymore, *i.e.* removing or adding one site from the analysis will not drastically change it. In order to test this belief, we classified the outliers according to their influence values, from the most negative to the least negative. We then deleted the i strongest outliers (for values of i ranging from 1 to 325) and inferred the ML-GTR tree. Using the Penny-Hendy distance (Penny and Hendy, 1985), we quantified the topological similarity of these 325 trees with each other and with the guide tree T. Penny-Hendy distance between two phylogenies calculates the minimal composition of elementary mutations which convert the first tree into the second one. From the dataset, we demonstrated that there were two stable trees. Removing any number between 2 and 44 of the strongest outliers led to almost the same tree. This is illustrated by the very small Penny-Hendy distance between these topologies (Fig. C.2). After removing the 46 strongest outliers an additional stable topology was found but the tree-to-tree logenies

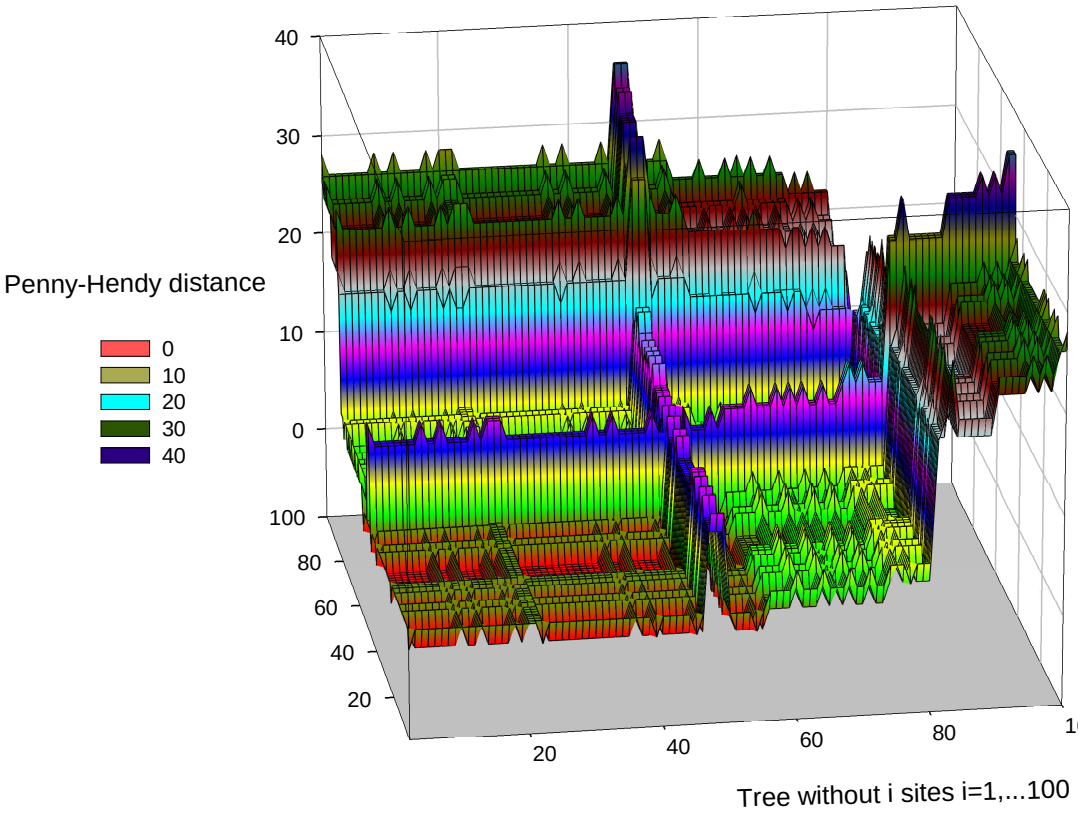

Figure C.2: Penny-Hendy tree-to-tree distances. x and y-axis figure the trees inferred after removing the i strongest outliers (i =1,. . . ,100). The ML-GTR guide tree (all data included *i.e.* $i = 0$) is not included on this figure.

We focus on the 325 sites with negative influence but we can probably concentrate on fewer sites. Huber (2004) proved the asymptotic normality of the influence value

under very general conditions. Using empirical mean and variance and given a Type I error level, it gives a practical solution to determine the threshold.

Strikingly, removing as few as the two strongest outliers already provides an improved stability: the majority of internal nodes in common with the ML tree have better bootstrap values (results ML-GTR and K2P-NJ tree reconstruction, data not shown). This further confirms the assumption that the removed information does not contribute to the ML tree.

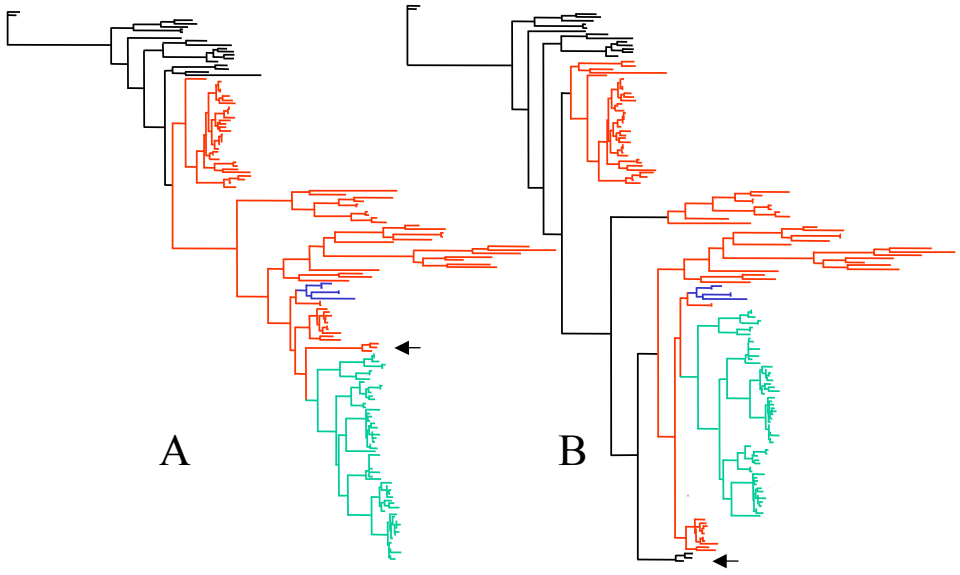

Figure C.3: (A) comparison of the tree topology for the ML-GTR guide tree vs. ML-GTR minus the strongest outlier and reciprocally for (B) the ML-GTR minus thestrongest outlier vs. ML-GTR guide tree. In black, part of the tree not affected by the removal of the strongest outlier. In green, phylum Glomeromycota. In red, phylum Zygomycota. In blue phylum Chytridiomycota. For the Zygomycota and Chytridiomycota notice that parts remained unchanged thus coloured in black. The arrow indicates the position of the 3 terminals *Mucor ramannianus Umbelopsis nana* and *Umbelopsis isabellina* as an example of a modification induced within the topology when removing the strongest outlier.

We take a closer look at the topologies inferred when removing the strongest outlier from the dataset to understand how and where it differs from the ML tree. Fig. C.3 shows the high magnitude of these differences. Different interpretations transpired from the trees inferred before and after removing the strongest outliers (Fig. C.3). First, the phylum Glomeromycota appeared remarkably stable and monophyletic. Only slight changes in the position of terminals can be detected when the trees generated from the dataset minus the strongest outlier to the dataset minus the 40 strongest were compared. These changes were observed within the cluster of 13 terminals containing 3 morphological species, *G. mosseae*, *G. claroideum*, and *G. lamellosum*. These changes might be attributable to the fact that these terminals are closely related and the quantity of molecular information was not high enough to clearly resolve their phylogenetic affinities. Second, the phylum Chytridiomycota appeared polyphyletic with a group of terminals containing *Basidiobolus* (two terminals), *Neocallimastix* (four terminals), one *Spizellomyces* one *Chytridium*, one *Pyromyces*, which was

weakly supported by bootstrap value (*i.e.* 51/55 respectively for MP and K2P-NJ) with the whole dataset, but the divergence of this group from the other Chytridiomycetes was reinforced when deleting the strongest outlier (bootstrap value = 63.5% and 66% respectively for MP and K2P-NJ), placed among terminals of the Zygomycota group. This result indicates that systematics within Chytridiomycota and Zygomycota must be re-evaluated, and this particular group must be re-classified within a Zygomycota sub-phylum. From these results we argue that the 2 Chytridiomycota groups have distinct evolutionary stories.

C.5 Acknowledgments

ABH and PV acknowledge "École thématique génomique environnementale" for helpful discussions. M. Bormans is acknowledged for comments on the manuscript. PV acknowledges GIS "Génomique marine" and "Fondation Total" and CNRS-ECCO for funding.

Appendix D

Taxon Influence

This section is a modified version of the article *Taxon Influence: Assessing Taxon-Induced Incongruities in Phylogenetic Inference* by M. Mariadassou, A. Bar-Hen and H. Kishino in revision for publication in Systematic Biology.

Abstract Understanding the evolutionary history of species is one of the heart of molecular evolution and is done using several inference methods. The critical issue is to quantify the uncertainty of the inference. The posterior probabilities in Bayesian phylogenetic inference and the bootstrap values in frequentist approaches measure the variability of the estimates due to the sampling of sites from genes and the sampling of genes from genomes. However, they do not measure the uncertainty due to the incorrectness of the statistical models of evolutionary processes. Species that experienced molecular homoplasy, recent selection, long branches and so forth may disrupt the inference and cause incongruences in the estimated phylogeny. We define a Taxon Influence Index to assess the influence of each species on the phylogeny. We found that although most species have a small influential, a small fraction of influential species strongly alter the phylogeny, even in clades only loosely related to them. We conclude that highly influential species should be given special attention and that removing such species from the dataset can lead to more reliable phylogenies.

Keywords: Bootstrap support, Taxon sampling, Taxon Influence, Tree Robustness

D.1 Introduction

The rapid increase in published genomic sequence for diverse organisms offers growing opportunities to infer the phylogenetic tree of groups of species. As with the estimate in any other inference problem, the inferred tree is subject to errors and uncertainties. Since many applications of phylogenetics require accurate phylogenetic estimates, it is crucial to determine how confident we can and should be in the inferred phylogenetic tree. Two main sources of uncertainty lie in variation among sites, studied in the bootstrap literature, and in variation among species, studied in the taxon sampling literature. The aim of this article is to quantify the influence of a taxon on the phylogenetic estimates.

The examples of rodents highlights the importance of good taxon sampling. Philippe (1997) and Cao et al. (1997) works on rodents show that introducing a few additional species in a phylogenetic study can have a strong impact on the inferred tree. In the rodent phylogeny studied in these two papers, claims of D'Erchia et al. (1996) that "the guinea pig is not a rodent" based on a 16-species phylogeny are seriously challenged when as few as 3 additional species are included in the analysis. In previous work, Lecointre et al. (1993) even argue that the number and choice of species included in the analysis has more impact on the inferred phylogeny than the choice of an evolutionary model. The field of taxon sampling has since been the focus of much attention (Hedtke et al., 2006; Hillis et al., 2003; Pollock et al., 2002; Zwickl and Hillis, 2002).

It is largely agreed on (Cao et al., 1994; Hedtke et al., 2006; Philippe, 1997; Poe, 2003; Poe and Swofford, 1999; Rannala et al., 1998; Zwickl and Hillis, 2002) that denser taxon sampling usually leads to more accurate phylogenies, especially for large number of species. Other studies (Pollock et al., 2002) also suggest that if an additional taxon is available, it is usually sound to use it in the inference before pruning it from the tree. However, the effect of an additional taxon depends on the position of this taxon

in the phylogeny (Geuten et al., 2007; Goldman, 1998); additional taxa that break long branches are expected to improve the stability of the tree (Heath et al., 2008), whereas adding additional long branches can hinder the stability and accuracy of the inference (Kim, 1998). It is also known that adding an outgroup can disrupt the ingroup topology even for small size topologies (Holland et al., 2003; Shavit et al., 2007). Furthermore, the yeast phylogeny studied by Rokas et al. (2003) and reanalyzed by Gatesy et al. (2007) shows that *removing* problematic taxa can lead to more stable and accurate phylogenies.

We introduce Taxon Influence Index (TII), a resampling method focused on the detection of highly influential taxa. TII quantifies the influence of a taxon on the phylogeny estimate by dropping it from the analysis and quantifying the resulting modifications of the inferred phylogeny. We also introduce the stability of a branch with respect to taxon sampling, defined as the number of taxa that can be dropped without altering the branch. We then use a real example (placental mammalian) to illustrate the utility of the method.

D.2 Material and Methods

D.2.1 Methods

Taxon Influence Index Let us consider an alignment of s homologous sequences of length n. Let $\mathbf{X} = (X_{pq})$ be the $s \times n$ matrix of data where X_{pq} denotes the state of the qth element in species p and is one of the 4 nucleotides or the 20 amino-acids. Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{iq})$ be the data at the ith row, or in other words, the sequence of the ith taxon.

Removing one taxon at the time from $\mathbf{X}$, we can generate s new smaller alignments: $\mathbf{X}^{(i)} = \mathbf{X} \backslash \mathbf{X}_i$. $\mathbf{X}^{(i)}$ is the alignment for all taxa but taxon i. Using any inference method (M), we infer T^* from $\mathbf{X}$ (the whole data set) and a smaller tree T_i from $\mathbf{X}^{(i)}$. We then prune taxon i from T^* to obtain T_i^* (Fig. D.1). We hereafter refer to T^* as the "whole tree", to T_i as the "inferred tree" and to T_i^* as the "pruned tree". The Taxon Influence Index (TII) of taxon i is the distance between trees T_i and T_i^*: $TII(i) = d(T_i, T_i^*)$.

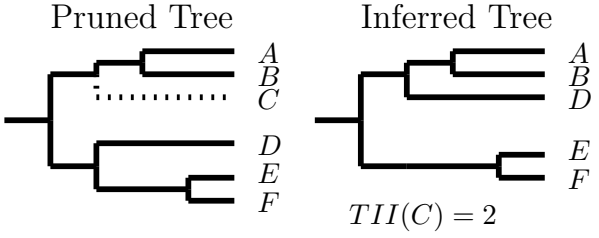

Figure D.1: Taxon Influence Index of taxon C. The pruned tree is obtained by pruning the taxon C from the complete tree. The inferred tree is inferred directly from $A, B, D{-}F$ only. The RF distance between the pruned and the inferred is 2 so $TII(C) = 2$.

The sequences are not realigned each time a taxon is pruned so that TII does not mix the influence of a taxon on the alignment and its influence on the phylogenetic

estimates. Whenever the tree is perfectly stable with regards to the taxon or when the influence of taxon i over the result is small, we expect the pruned tree and the inferred tree to be very similar. In the extreme case where a taxon is duplicated in the alignment, the two copies of this taxon should be clustered at a tip of the topology and removing any of them should not modify the topology, nor the tree, in the slightest. Although TII can be used with any of the several distances on trees, we focus here on two distances: Robinson-Foulds (RF) distance (Robinson and Foulds, 1979) and Branch Score Distance (BSD) (Kuhner and Felsenstein, 1994). RF accounts only for topological differences whereas BSD weights the topological differences by the length of the affected branches.

Branch Stability TII can detect any influential taxon but the pattern of the changes is also interesting. For example, are the branches affected when removing a taxon always the same (indicating some weakness of these particular branches) or are they well distributed across the tree (indicating some weakness of the inference method)? The study of branch stability is helpful to answer these questions; internal branches of the tree are scored for their robustness to taxon removal and more generally, changes in the taxon sample. A branch not affected by taxon sampling is robust and can reasonably be trusted whereas a branch affected by many species, even some far away from it, is highly sensitive to taxon sampling and should be considered cautiously.

We define the branch stability $BS(b)$ of an inner branch b (of the whole tree) as the number of pruned trees T_i in which it is also present. Formally $BS(b) = \sum_{i=1}^{s} \mathbb{1}_{T_i}(b)$ with $\mathbb{1}_{T_i}(b)$ being 1 if b is present in T_i and 0 otherwise. Since T^* has a one more leaf than T_i, it also has one more inner branch and so not all branches of T^* have a counterpart in T_i. Indeed, an inner branch connected to the terminal branch of i disappears when pruning i. Since topologies are binary trees, inner branches can be connected to either 0, 1 or 2 terminal branches and so there are three possible maximum values for BS scores: $s-2$, $s-1$ and s. Hereafter and for easier comparison with usual support values, the $BS(b)$ values are expressed as a percentage of their maximum value and range in $[0, 100]$.

D.2.2 Material

Sequence Data We examined an empirical data set taken from Kitazoe et al. (2007) consisting of mitochondrial protein sequences (3658 amino acid sites in total) from 61 placental mammals, belonging to *Laurasiathera*, *Supraprimates*, *Xenartha* and *Afrotheria* plus seven outgroup taxa, belonging to *Monotremata* and *Marsupialia*. The gaps are excluded and the sequences are not realigned when removing a taxon. Although the original data set contains 69 species, two of them, labelled tenrec1 and tenrec2 are genetically so close, as shown by the phylogenies published in Kitazoe et al. (2007), that we decided to keep only tenrec1 and relabelled it tenrec. Our placental mammal data set thus consists of only 68 species, instead of 69 in the original data. As pointed out by Kitazoe *et al.*, these data present relatively long sequences, good taxon sampling and very little missing data. Another advantage of mammals is that their phylogeny has been intensively studied and that many problems and hard-to-resolve

clades have been identified (Prasad et al., 2008). Of particular interest is the position of the guinea pig (*Cavia porcellus*) in the order *Rodentia*, which has long been a heated issue among molecular phylogeneticists (Belfiore et al., 2008; Cao et al., 1994, 1997; D'Erchia et al., 1996; Graur et al., 1991; Hasegawa and Fujiwara, 1993; Philippe, 1997).

D.2.3 Phylogenetic Analysis

While TII is amenable to phylogenetic inference method, we restricted the analysis to Maximum-Likelihood (ML) for the sake of brevity.

Evolution Model Phylogenetic trees were inferred using PhyML 2.4.4 (Guindon and Gascuel, 2003). We used the mtMam+I+Γ4 model, selected by ProtTest 1.4 (Abascal et al., 2005) as the best model no matter what the criterion (AIC, BIC, etc). The mtMam empirical rate matrix is the one used in the best four models (mtMAM and any combination of +I and +Γ4), followed by mtREV in the next four models. The hill-climbing search was initiated at the BIONJ tree of the alignment, the default starting tree for PhyML. We used PhyML to optimize the branch lengths, the α parameter of the Γ shape and the proportion of invariable sites (command line: `phyml alignment 1 s 68 0 mtMAM e 4 e BIONJ y y`). Thanks to the moderate size of the data set (68 species, 3658 AA), each of the 69 trees (1 for the whole data set and 1 for pruning each of the 68 species) was inferred in ~ 45 minutes CPU time (on a 1.66-GHz Intel Core Duo PC). We also performed 200 replicate ML bootstrap analyses for the whole data set (68 species) in a bit more that 3 CPU days.

Inference Software Since topologies are of great importance in computing species leverage, we also inferred ML topologies using RAxML 7.0.4 (Stamatakis, 2006) and compared them to the PhyML topologies on the whole data set and on a few smaller data sets. The model used in both cases was the same: mtMam+Γ4 (command line: `raxmlHPC -s alignment -n output_name -m PROTGAMMAMTMAM -e 0.01`, the `-e` option was used to compute the likelihood with a precision of 0.01 instead of the default 1). Highest likelihood trees from multiple runs of RAxML were the same as highest likelihood trees from multiple runs of PhyML for most of the data sets tested and very similar for the others. We used PhyML because a single run was roughly 4 times faster than RAxML (~ 45 min against ~ 3 hours).

Analyses with PhyML were scripted using custom shell scripts. The TII and BS scores were computed using *R* scripts (available on demand from ABH).

D.3 Results

D.3.1 Inference Quality

As we use ML, the inferred tree T_i should have a higher likelihood than T_i^*. After all, T_i is inferred to maximize the likelihood over $\mathbf{X}^{(i)}$ whereas T_i^* maximizes the

likelihood over **X** before being pruned. Although we expect the likelihood values of T_i and T_i^* to be close and even to correlate quite well as only a fraction of the taxa strongly affect the inference, they should systematically be higher for T_i than for T_i^*. Results from our analyses (data not shown) confirm it.

D.3.2 TII Distribution and Influential Taxa

SLI values of the species are plotted in Figure D.2. We note that guinea pig has the highest TII (12), confirming previous findings of guinea pig being a rogue taxon. The result is robust to model choice (with or without Rate Across Sites (RAS), and with mtREV instead of mtMAM), with guinea pig TII always being the highest, between 12 and 14. The comparison of the pruned and inferred tree (not shown) for guinea pig reveals that removing as little as one species can affect the topology even in remote places; removing the guinea pig disrupts the clades of the insectivores and modifies the position of the northern tree shrew (*Tupaia belangeri*), 6 branches away from it.

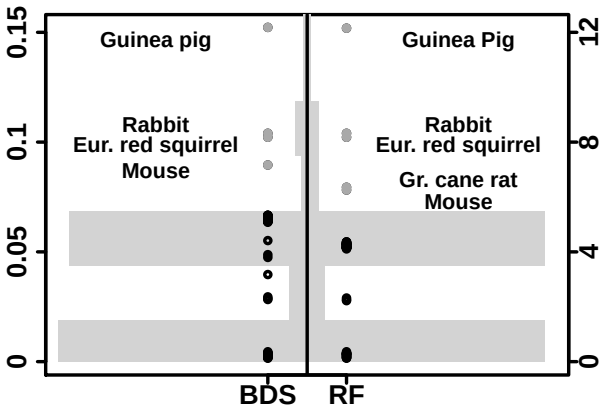

Figure D.2: Dot-plot and histogram of *TII* values for BSD (left) and RF (right) distance. Taxa with TII higher ≥ 0.75 (BSD) or ≥ 8 (RF) are labelled with their name. Though many taxa have the same TII, and thus should have exactly the same position in the dot plot, the corresponding superimposed points are slightly shifted for better legibility, .

Using a cutoff value of 8, which represents two standard deviations from the mean, three species are identified as influential (marked in bold and annotated in Fig. D.3) and concentrated among *Glires*: guinea pig, European red squirrel (*Sciurus vulgaris*) and rabbit (*Oryctolagus cuniculus*). No matter what distance is used (RF or BSD) the same species stand out as influential (Fig. D.2) and the TII-induced order is conserved; only 4 of the remaining 65 species change rank when changing the distance. But the number of influential species is highly dependent on the model: it varies from 4 for the mtMam+I+Γ to 10 $\sim$ 12 for mtMam+Γ and mtREV+Γ. Fortunately, there is important overlap; for example, the 3 species influential under mtMam+I+Γ are part of the set of species influential under mtMam+Γ. Conversely, 20 species (again varying with the model from 7 in mtREV/mtMAM+Γ to 20 in mtMam+I+Γ), in bold in Figure D.3, are extremely stable in the sense that their removal does not disturb the topology at all. Remarkably, the stable species are well distributed over the tree and either part of a clade of two species or at the end of a long branch.

D.3.3 Branch Stability

With the exception of influential species and extremely stable species, most of the *TII* values are 4. This means that most inferred trees are slightly different from the corresponding pruned trees, with a difference of only two branches. We use the stability scores to check whether these differences are well distributed over the whole topology T^* or concentrated on a limited number of branches. The results are shown in Figure D.3 (inset). Interestingly there is no correlation between BS scores and branch lengths (data not shown) even when restricting the analysis to the branches with BS < 100%.

Two branches with very low BS scores belong to the *Afrotheria* (Fig. D.3), indicating a poorly resolved clade. Indeed, even if a species is only weakly connected to the *Afrotheria*, removing it from the analysis often changes the inner configuration of the *Afrotheria* clade. These branches have very low bootstrap values (from the root to the leaves 45%, 54%). A detailed comparison between BS scores and bootstrap values is informative about their similarities and differences. First, bootstrap is more conservative than BS: all branches with 100% bootstrap values also have 100% BS, but some branches (20) with 100% BS do not have 100% bootstrap values (marked in light grey in Fig. D.3). Second, for the 9 branches whose both BS and bootstrap values are lower than 100% (marked in dark grey in Fig. D.3), the correlation is very low (0.25). Except for the two branches aforementioned, the bootstrap values are much smaller than their BS equivalent: they vary between 11% and 75% whereas all BS scores are over 92%.

D.4 Discussion

D.4.1 Influential Taxa and Rogue Taxa

TII is used to detect influential species, that is to say species which strongly impact the phylogenetic estimates. TII shares similarity with Lanyon (1985) approach but the focus is different; Lanyon aimed at detecting incongruences in the data and building a consensus whereas we focus on detecting influential species. TII is closer in spirit to Thorley and Wilkinson (1999) measure of leaf stability. However, leaf stability examines only the impact of taxon on triplets whereas TII examines its impact on the complete topology. Our procedures also shares resemblance with Rosenberg and Kumar (2001) and Pollock et al. (2002) but both authors are interested in the general impact of taxon sampling on the overall accuracy and thus consider a simulation experiment in which the true phylogeny is known, which is often not the case in practice. TII is more general as it both quantifies the influence of each species and the stability of each branch to taxon sampling.

Finally, an influential taxon might not be a rogue taxon. The term “rogue” is generally restricted to taxa whose presence impedes phylogeny estimation (Sullivan and Swofford, 1997; Wilkinson, 1996) and the characterization of a taxon as such requires further investigations and independent lines of evidence. Indeed, a taxon which has a stabilizing, beneficial effect on the phylogeny estimate is certainly influential but

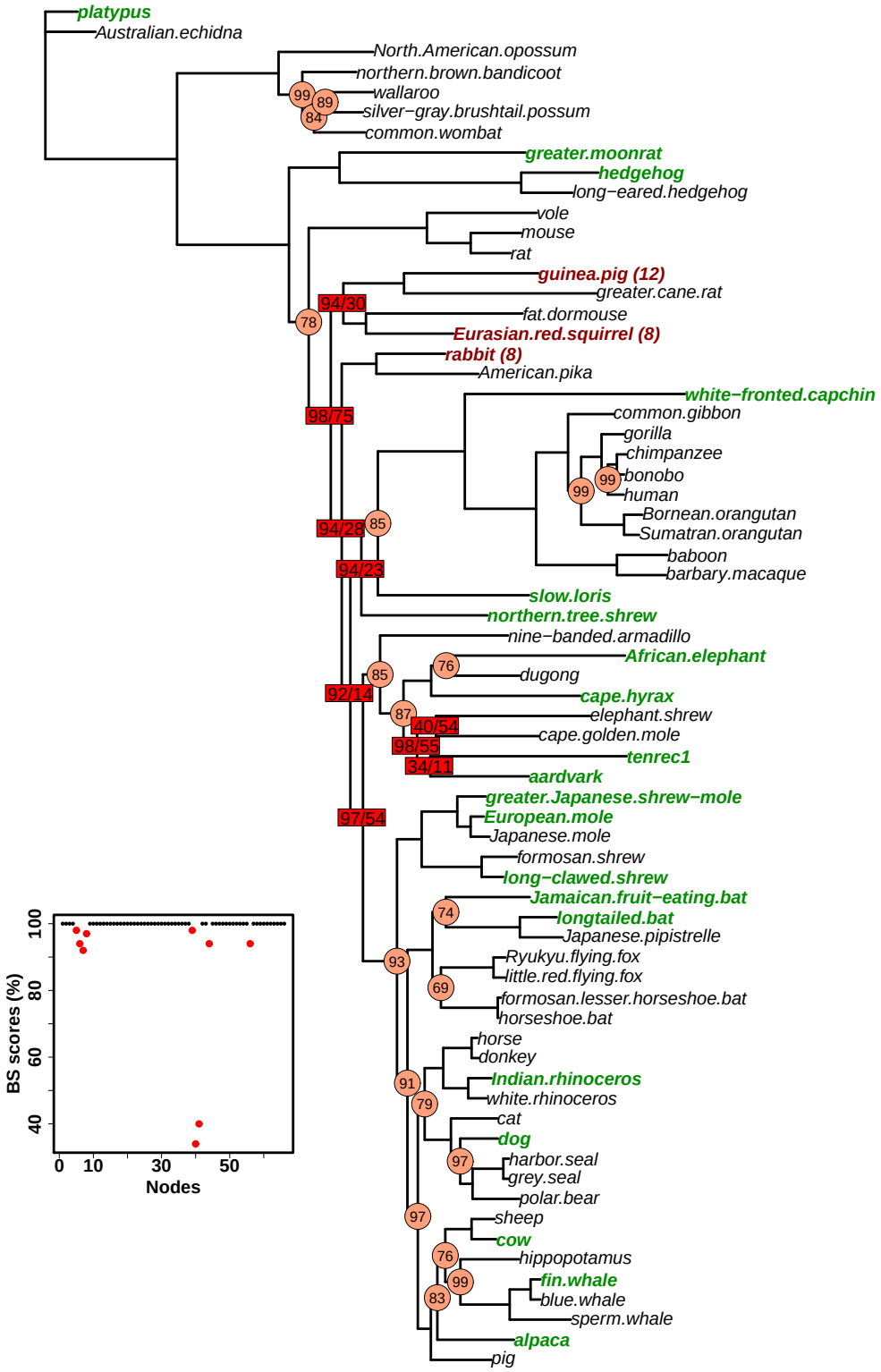

Figure D.3: Placental mammals phylogeny with BS scores and bootstrap values. 100% bootstrap and BS values are omitted. Light grey branches have 100% BS scores but <100% bootstrap (circles), whereas dark grey branches have <100% BS and bootstrap scores (resp. left and right in the rectangle). Influential taxa (TII>7) are in bold and annotated by their TII, stable taxa are in bold. Inset: BS scores (in %) of internal branches.

definitely not rogue. Therefore influential taxa should not be discarded from the analysis but rather given special attention. Discovering why only some taxa and not some others disrupts the phylogeny is better used to understand how and why evolution model fail us, or to try a denser taxon sampling of the putative group of influential taxa.

D.4.2 TII and BS scores

TII and BS scores are strongly connected since:

$$\sum_{i \in \text{species}} TII(i) = \sum_{b \in \text{branches}} (BS_{max}(b) - BS(b))$$

Trends in BS scores propagate to TII values and vice versa: a high average TII means a low average BS score and vice-versa. In the placental mammal phylogeny, only 20 taxa have absolutely no impact on the tree topology when pruned from the dataset. This fraction is small at first sight but it is artificially so because of the two unstable clades: the first one consisting of aardvark (*Orycteropus afer*) and tenrec (*Echinops telfairi*), the second one of elephant shrew (*Elephantulus sp. VB001*) and cape golden mole (*Chrysochloris asiatica*). These two clades account by themselves for 37 taxa with a TII of 4. The number of taxa modifying the tree elsewhere than in these two branches reduces to 11: American pika (*Ochotona collaris*), cape golden mole, dugong (*Dugong dugon*), elephant shrew, Eurasian red squirrel, fat dormouse (*Myoxus glis*), greater cane rat (*Thryonomys swinderianus*), guinea pig, mouse (*Mus musculus*), nine-banded armadillo (*Dasypus novemcinctus*) and rabbit. As expected, most taxa leave the tree completely or almost completely unchanged. Half of the stable taxa belong to clades with only two species. This is not surprising because the two species of such a clade, especially if the terminal branches are very short, have very similar sequences and are almost redundant. Removing any one of them affects the inference process only marginally. For sister taxa with short terminal branch lengths, it might be worthwhile to prune the two taxa at the same time.

D.4.3 TII and Long Branches

When two non-adjacent taxa share many homoplasic character states along long branches, some methods (most famously parsimony) interprets such similarity as homology. The resulting tree depicts the two taxa as sister to one another, attributing the shared changes to a branch joining them; this effect is termed long-branch attraction (LBA) (Felsenstein, 1978). We can therefore expect taxa at the end of long terminal branches to affect the inference and have high TII.

And indeed, the 11 remaining species with positive TII after controlling for the two unstable branches are at the end of terminal branches significantly longer than the average terminal branch (Wilcoxon signed ranks test, increase = 81%, $p = 0.002$). However, the reverse is not true, only 8 (44%) from the 17 taxa at the end of the 25% longest terminal branch lengths are influential. And this ratio never exceeds 47%, achieved for the 20% longest terminal branches. Influential taxa are therefore not just an artifact of long terminal branches.

D.4.4 Relation with bootstrap support

The most popular method to assess uncertainty is to compute bootstrap values. This approach has strong theoretical justification in certain circumstances (Hasegawa and Kishino, 1989; Kishino and Hasegawa, 1989; Shimodaira and Hasegawa, 1999) but the link between bootstrap values and support for a clade is far from straightforward (Hillis and Bull, 1993; Zharkikh and Li, 1992). More importantly, bootstrap aggregates different sources of uncertainties and is unable to pinpoint specific sources of uncertainty, be it problematic sites (Bar-Hen et al., 2008), taxa or branches. Finally, bootstrap was never intended to study the uncertainty induced by taxon sampling.

Most branches are highly resilient to taxon sampling and only a few, poorly resolved to begin with are clearly affected by taxon sampling. Comparison of BS scores to bootstrap values suggests that taxon sampling makes a relatively small contribution to phylogenetic variability compared to site sampling. More importantly, branches with BS scores <100% are also among those with lowest bootstrap values. The very low bootstrap values of these branches probably arise from two correlated causes. First, the branch might just be wrong or intrisically hard to resolve because it encompasses taxa whose positions in the tree are unclear. These taxa, by being in poorly sampled clades, or by exhibiting peculiar features, are not easily placeable and could be at several places in the tree so that several topologies are almost equally likely and the branch of interest will not be in all of them. There is no real phylogenetic signal supporting these branches and any subtle modification of the data set, be it pruning a taxon or bootstrapping sites will result in a different topology. Bootstrapping sites modifies the alignment to a greater extent than pruning a taxon and mimics the stochastic variations induced by sampling the sites. It thus captures at least two sources of variations for these branches: on the one hand, excessive sensitivity to the alignment induced by influential taxa and on the other hand normal sensitivity to the size of the alignment, predicted by standard sampling theory. BS scores help isolating the two sources.

D.4.5 Extension to Bayesian methods

TII is amenable to any inference method but requires the output to be a single tree. Even tough the result of bayesian phylogenetic inference is often summarized as a majority-rule consensus tree (MRC) (e.g. MrBayes, (Ronquist and Huelsenbeck, 2003)) one of the strengths of bayesian methods is the ability to account for uncertainty by providing a posterior distribution of the topology instead of a point estimate. Following Cranston and Rannala (2007) approach on agreement subtree, TII and BS scores can be modified to use more of the posterior distribution than just the MRC tree.

As for BS scores, the modification only consists of weighting each branch by its posterior probability $Q_i(b)$ in each of the pruned tree: $BS(b) = \sum_i Q_i(b)$. $Q_i(b)$, as a posterior probability, can take any value between 0 and 1. The ML equivalent would be to weight each branch b by its bootstrap proportions, instead of 1 if the branch is recovered and 0 else, as we did.

To compute TII, we need to measure the "distance" between posterior distribu-

tions, instead of between trees. Although many such distances exist (Gibbs and Su, 2002), we believe they are not well-adapted to this problem; none of them includes any information about the tree structure. We instead propose the following, inspired by Cranston and Rannala (2007). We build, for each species i a pruned posterior distribution Q_i^* from the complete posterior distribution Q^* by pruning species i from the topologies of the posterior and condensing resulting identical trees. In addition, we also compute an inferred posterior distribution Q_i (from the small data set $\mathbf{X}^{(i)}$). The TII is then defined as the average distance between a random tree from Q_i^* and a random tree from Q_i,

$$TII(i) = \mathbb{E}_{Q_i^* \otimes Q_i}[d(T, T')] = \sum_{T,T'} Q_i^*(T) Q_i(T') d(T, T')$$

with $d(.,.)$ is the same distance as before. Although with this definition the TII is not strictly a distance anymore, it is easy to compute as the average distance between the MCMC of Q^* and Q_i once they reach convergence.

D.4.6 Limitations and future work

TII and BS scores computations require the estimation of quite a few trees; the number of trees to infer grows linearly with the number of taxa, and because of increasing complexity with a larger number of taxa, the inference time for each of them of course increases. The last point is not specific to the proposed measures and also holds for many quantities computed on trees. The total computation time increases more than linearly with the number of species. Computation of TII values and BS scores is fast as it only requires comparison at the branch level. TII is thus useful for moderate datasets but not for very large ones.

By pruning only one taxon at the time, we are able to detect single taxon that exhibit peculiar evolutionary features, as corroborated by previous findings about the guinea pig (Cao et al., 1997), but we are unable to detect troublesome groups of taxa. To do so, we would need to remove two, three or more taxa at a time. The large number of possibilities make inference of all the small trees unrealistic. The most promising paths to tackling this problem are to cluster the taxa or to remove them sequentially. The first option is to "cartoon" the phylogenies by clustering taxa in groups whose inner phylogeny is well supported and choosing one representative in each group while discarding all the others to reduce the size of the topology. The second option is to remove the taxa sequentially: remove the highest TII taxon first, compute the TII again on the remaining taxa, remove the new highest TII taxon and so on until either a given number of taxa have been filtered or the highest TII does not exceed some threshold. The results so far in this direction (not shown) suggest that the taxa emerging as influential when removing one taxon, then two and so on are always the same. The branches of the tree with low support also are the same, even after removing a few species, confirming that they are intrinsically hard to resolve.

Bootstrap and posterior probabilities are good ways to assess the uncertainty induced by site sampling but aggregate many sources of uncertainty with no way to easily ascertain which factors contribute most. They are also are much more difficult to correctly interpret than thought at first (Yang, 2007). Furthermore, they were never

intended to study the impact of taxon sampling on the inference. We show that some taxa have a large impact on the phylogenetic estimation and propose an index to identify them quantitatively with the ambition to better characterize the factors of uncertainty in phylogenetic reconstructions.

D.5 Acknowledgements

We are indebted to L. de Oliveira Martins, Z. Yang, P. Vandenkoornhuyse and J. Felsenstein for useful discussions and comments on earlier versions of the manuscript. We would also like to thank J. Sullivan, C. Ane and an anonymous reviewer for reviewing the manuscript and for their insightful comments. This research was supported by a grant from the Collège Doctoral Franco-Japonais.

Chapter 6

Discussion and Prospects

6.1 Summary

In Part I, we were interested in various ways to control the variance of the likelihood score of a tree in a parametric framework. In Part II, we forget this goal and rather focus on non-parametric methods to detect outliers in the alignment, be it sites (Chapter C) or taxa (Chapter D), which potentially hinder the variance control. The two methods presented here are very similar in their execution, jackknifing sites or species, but have very different justification.

Jackknifing sites Sites are (at least assumed to be) independent and identically distributed. They contribute to the variability of the phylogenetic estimates through the sampling variability. Resampling techniques, such as bootstrap and jackknife, are legitimate for such i.i.d. random variables. They can be used to estimate the variability of estimates, either directly (bootstrap) or through extrapolation (jackknife, see e.g. Huber (1981)). We use jackknife differently, not to infer the variability of the estimates but to compute the influence of each site on the estimates, via influence curve. Influence curves work best for estimates valued in $\mathbb{R}^d$, or at least in a metric space $(\mathcal{X}, d)$. Unfortunately, inferring a phylogeny requires estimating several parameters of different natures (discrete/continuous). We typically estimate a topology (discrete), the associated branch lengths (continuous) and parameters (continuous) of the evolution model.

We could compute the influence of each site on each parameter, but this approach is only meaningful for a subset of the parameters. Parameters of the model of sequence evolution can be considered independently of others parameters. Indeed, the Markov chain used to model the characters is often assumed to have reached its stationary distribution so that the model of sequence evolution is specified independently of the tree; we can compare parameters of the model before and after jackknifing a site with no knowledge of the tree. If the chain is not stationary (Boussau and Gouy, 2006), we need some information about the initial distribution and the total length of the tree to compare the models. Branch lengths are in a similar situation; comparing the branch lengths before and after jackknifing a site is meaningful only if they correspond to the same branch. Branch lengths must be considered jointly with the topology.

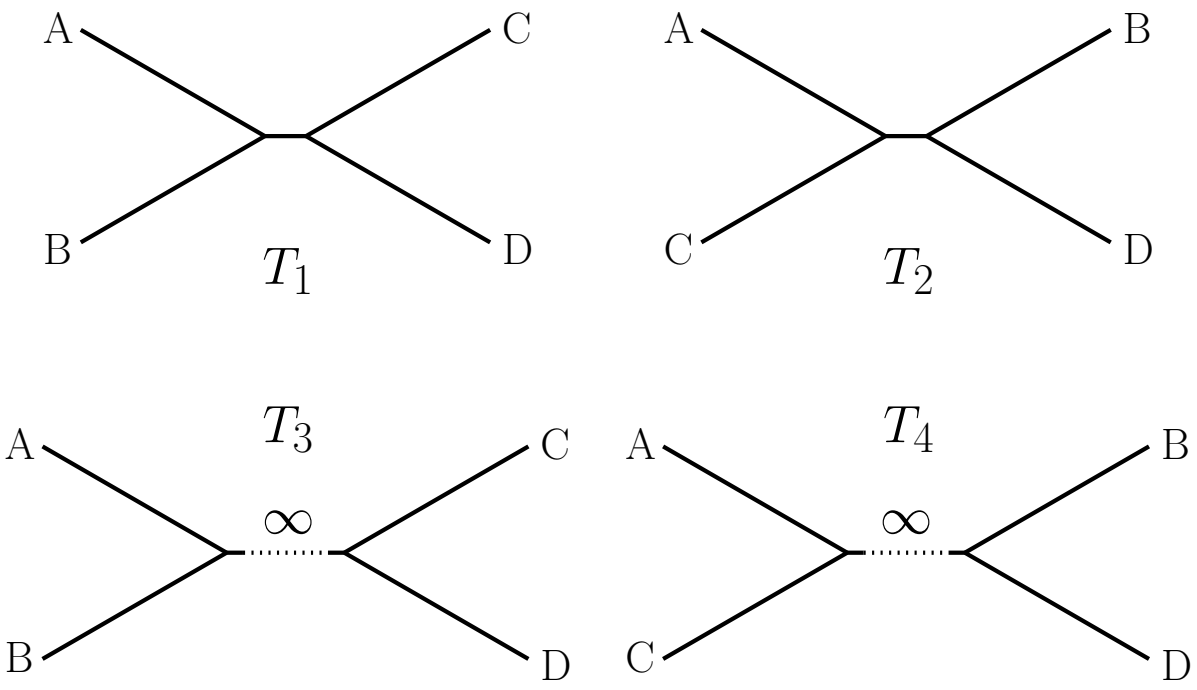

Figure 6.1: Top: Trees T_1 and T_2, terminal branch lengths are 0.1, central branch length is 0.02. Bottom: Trees T_3 and T_4, terminal branch lengths are 0.1, central branch length is ∞. $d_{RF}(T_1, T_2) = d_{RF}(T_3, T_4) = 2$ but T_1 and T_2 are significantly closer than T_3 and T_4.

As for the topology, although it is possible to combine topology and branch lengths to embed them in the so-called *BHV* metric space (Billera et al., 2001), the metric of the *BHV* involves finding geodesics in a complex space (made of $(2s-3)!!$ positive orthant $[0,\infty)^{s-2}$ glued together at their boundaries) and is unfit for practical computations. Robinson-Foulds (RF) distance simply counts the number of nodes which differ from one topology to another. Other tree metrics such as Subtree Pruning and Regrafting (SPR), Nearest-Neighbor Interchange (NNI) or Tree Bisection and Reconnection (TBR) count the minimum number of elementary moves (Sec. 2.3.3) required to to go from one topology to another. RF is easy to understand and to compute and gives a simple insight as to what extent the topology is changed when jackknifing a site. But RF shares with the SPR, TBR the NNI the drawback of being discrete and not concerned with branch lengths. The latter drawback means that changes affecting a small branch are given the same importance as changes affecting a medium to long branch (see Fig. 6.1), which can be justified by the sake of simplicity or else but at least needs to be justified.

We instead adopted an original estimate: the log-likelihood $S(F_n) = \frac{1}{n}\log P(X_i; \hat{T}, \hat{M})$ of the ML parameters. In our opinion, the first benefit of this statistic is to aggregate all parameters, albeit in a non standard way. The second benefit of likelihood score, compared to other estimates, is its simplicity: a single value, instead of a topology or a list of values. Finally the sign of $S(F_n) - S(F^{(h)}_{n-1})$ gives qualitative information about site X_h. $S(F_n)$ is an unbiased estimator of the log-likelihood of a site. The higher this value, the closer $P(.; \hat{T}, \hat{M})$ is to the empirical distribution. In other words, high values of $S(F_n)$ correspond to well supported trees. Therefore, positive values of $S(F_n) - S(F^{(h)}_{n-1})$ reflect that the tree computed on the whole alignment is better supported than the tree computed on the whole alignment but X_h and

vice-versa for negative values of $S(F_n) - S(F_{n-1}^{(h)})$. This intuitive interpretation is appealing. Jackknifing of sites is amenable to any inference method and adaptation to non-likelihood methods if of course possible. For parsimony inference for example, we would consider the parsimony score per site instead of $S(F_n)$.

Jackknifing species Unlike sites, species are not independent and identically distributed. Instead, they have a structure, called phylogeny, that we try to discover and are highly clustered on this phylogeny. It is thus not easy to see at first sight what statistical meaning there is to jackknifing species. However, we do not advocate the use of jackknifing species as a theoretically well-founded statistical tool to infer the variability of an estimate but rather as a useful exploratory data analysis tool to detect influent species and potential bias.

Taxon Influent Index (TII) use jackknife across species, removing one species from the data set to see the effect of pruning a terminal branch on the estimate of the rest of the tree. Our prior belief was that pruning a taxon should not, or only locally, change the new estimate. After all, the only exclusive information about the evolution process brought by a new species corresponds to its terminal branch; the rest is duplicated in other species of the alignment. This prior belief was nevertheless contradicted by study of the placental mammals phylogeny. Some species, previously identified in the literature as rogue species and singled out by their extreme TII values, hinder the inference, even in deep nodes.

Taxon sampling is far from being an exact science and while some rules of the thumb (denser taxon sampling leads to more accurate phylogenies, denser taxon sampling is more beneficial when cutting long branches, etc) seem pretty efficient (Heath et al., 2008; Hedtke et al., 2006; Lecointre et al., 1993; Poe, 2003; Poe and Swofford, 1999; Pollock et al., 2002; Rannala et al., 1998; Zwickl and Hillis, 2002), adequately sampling taxa is more art than science. Taxon sampling is subject to availability constraints and arbitrary choices (choice of an outgroup, choice of species to include in the analysis). It is thus in our opinion essential to assess the extent to which a tree is vulnerable to a few taxa.

Unlike site outliers detection, log-likelihood comparison is not an option. There are at least threes issues with likelihood scores when pruning a species from a tree. The first one is that TII is amenable to any inference method, many of which do not do not allow for computation of the log-likelihood of the estimates. The second one is that the compared trees correspond to different models, not nested in each other (as reflected by the different number of terminal branches). The bigger tree correspond to a more complex model and its log-likelihood is bound to be lower than that of the simpler model. Binary variables are helpful to understand the problem. Consider $(X_1, Y_1), \ldots, (X_n, Y_n)$ a series of i.i.d. couples of binary variables. Imagine the distribution of (X_1, Y_1) is $P_{p_x,p_y} = \mathcal{B}(p_x) \otimes \mathcal{B}(p_y)$. The ML-estimator of (p_x, p_y) is $(\bar{X}_n, \bar{Y}_n)$ and the log-likelihood score would be:

$$\begin{aligned}\frac{1}{n}\log P_{\bar{X}_n,\bar{Y}_n}\Big((X_1,Y_1),\ldots,(X_n,Y_n)\Big) &= \frac{1}{n}\log\left(\mathrm{C}_n^{n\bar{X}_n}\right) + \bar{X}_n\log\bar{X}_n + (1-\bar{X}_n)\log(1-\bar{X}_n)\\ &\quad + \frac{1}{n}\log\left(\mathrm{C}_n^{n\bar{Y}_n}\right) + \bar{Y}_n\log\bar{Y}_n + (1-\bar{Y}_n)\log(1-\bar{Y}_n).\end{aligned}$$

When interested only in the first of the two variables $((X_i))$, the ML estimator of p_x

still is $\bar{X}_n$ but the log-likelihood score becomes:

$$\frac{1}{n}\log P_{\bar{X}_n}(X_1,\ldots,X_n) = \frac{1}{n}\log\left(C_n^{n\bar{X}_n}\right) + \bar{X}_n\log\bar{X}_n + (1-\bar{X}_n)\log(1-\bar{X}_n).$$

which is $-\frac{1}{n}\log C_n^{n\bar{Y}_n}\bar{Y}_n^{n\bar{Y}_n}(1-\bar{Y}_n)^{n(1-\bar{Y}_n)}$ higher simply because the first model deals with higher dimensional data and has a flatter likelihood landscape. In this simple example, the estimate of p_x is of course the same in both model. Comparison of the likelihood scores requires preliminary correction. Here the variables are simple enough and the correction is pretty straightforward, but it might be more complex if there was unknown correlation between X and Y.

As proposed in Chapter D, we can always prune species i from the whole tree T^* to obtain the prune tree T_i^* and compare it to the inferred tree T_i. T_i^* and T_i have the same leaves and we can thus compare the log-likelihoods of T_i^* and T_i. But T_i is the ML estimate of the pruned data set; its log-likelihood score is bound to be better than any other tree, including T_i^*. The log-likelihood difference is always positive and is not as easily interpretable as for influence functions.

Our need for a simple, easy to understand influence index led us to use the RF-distance (or some other distance) between T_i and T_i^*. It is a partial, flawed index. As discussed in previous paragraph, it does account only for changes in the topology, and at best in the branch lengths, but not in parameters of the model of sequence evolution. Still, TII only pretends at detecting influent taxa and it was powerful enough with RF-distance to detect influent taxa (actually rogue ones) in the placental mammal phylogeny.

6.2 Further Work

Studying the resistance of phylogenetic estimates to small perturbations of the data (bootstrap resampling, jackknifing of sites, jackknifing of species) is a promising field. We pointed out in Section 5.2 some of the flaws and drawbacks of bootstrap. We are fully aware that IF and TII are also potentially flawed. We list here some of the challenges raised by the use of IF and TII and promising ways to solve them.

Setting a threshold Computations of IF and TII raise at least one important question: at what value does a site qualify for outlier or a species for rogue ? In chapters C and D, we evaded the problem by focusing only at the most extreme values, which were well separated from the bulk of other values. It would however be more desirable to compare all values, instead of only the most extremes, to a threshold in order to detect influent sites and label them as such. Calibrating the threshold requires some knowledge about the behavior of TII and IF.

Results on influence function and the sensitivity curve (Cuevas and Romo, 1995; Nowak and Bar-Hen, 2005) suggests that IF values can be used to estimate the variance of the log-likelihood $S(F_n)$ via:

$$\widehat{\mathrm{Var}\left(S(F_n)\right)} = \frac{1}{n}\sum_{h=1}^{n} IF_{S,F_n}(X_h)^2.$$

Although the distribution of the $IF_{S,F_n}(X_h)$ remains unknown and thus no confidence interval can be properly calibrated, $\widehat{\text{Var}\left(S(F_n)\right)}$ provides a useful scale to assess the exceptionality of a $IF_{S,F_n}(X_h)$ and trim the outliers deemed too influential.

Unfortunately, The literature provides no similar result for TII values. Nevertheless, detecting influent species only require knowledge of a "typical" variation scale induced by small modifications of the data. This typical scale can then be compared to TII values to decide which species are influent. This does not completely remove the arbitrary from the decision to label a species as influent but at least reduces it. There are many ways to create scales to which TII can be compared. The most obvious scale is of course the variation scale induced by jackknifing species and a candidate threshold would be two standard deviations above the average ($\mu + 2\sigma$) by Gaussian analogy. Parametric bootstrap could also be used to calibrate the threshold. However, we think that other scales are more relevant. We need to compare TII values to an independent source of variability, such as site sampling. After all, if site sampling induces ten times the variability induced by taxon sampling on the topology, we should focus on sequencing longer sequences before worrying about adequate taxon sampling. A taxon is influent, and potentially rogue, only insofar as it is able to override the noise induced by site sampling (and other sources of error) and to pull the inferred topology in one direction. Whether to use the scale provided by variation in the topology when bootstrapping sites, jackknifing them or some other perturbations has not been solved yet. Note that the selected scale can also be used for *IF* values.

Set of Trees Another issue raised by, but not limited to, *IF* and *TII* computations is our belief that a point estimate, here a single tree, is enough to capture the variability induced by a site or a species. Imagine for example that two trees $T*$ and $\widetilde{T}*$ come out as the ML tree of the whole alignment and two trees T_i and $\widetilde{T}_i$ come out as the ML tree of the alignment missing species i. There are then 4 different ways to calculate $TII(i)$: $d(T_i^*, T_i)$, $d(\widetilde{T}_i^*, T_i)$, $d(T_i^*, \widetilde{T}_i)$ and $d(\widetilde{T}_i^*, \widetilde{T}_i)$, each of which may yield a different value. It is unclear whether we should pick one of the different combinations, in which case how, or try to combine them is some. Of course, the event that two trees have exactly the same likelihood score and turn out to be the ML tree is extremely unlikely but softer versions are much more frequent. Instead of having the exact same likelihood score, T^* and $\widetilde{T}^*$ just need to have likelihood scores so close they are not significantly different, as assessed by the test of topology presented in Section 2.4.3. *TII* values can in principle be calculated on any pair of tree $(\widetilde{T}^*, \widetilde{T}_i)$ given that $\widetilde{T}^*$ is not significantly different from T^* and $\widetilde{T}_i$ from T_i and span a wide range of value.

IF are computed from log-likelihood scores and hence, immune to some extent to this problem. Even if several trees with very different topologies lie only a fraction of likelihood unit from the ML tree, their likelihood scores are by construction very close. No mechanism prevents *TII* to vary widely when choosing different trees in the set. This criticism holds for any quantity computed on the ML tree without accounting for uncertainty in the tree and is certainly not limited to IF and TII. One way to address this problem is to consider sets of trees rather than just the ML tree. TII values can then be averaged over all possible pairs. Formally, if $(T_i^{*(k)})_{k=1..K}$ is the set of trees replacing T_i^* and $(T_i^{(l)})_{l=1..L}$ the set of trees replacing T_i, TII changes from

$TII(i) = d(T_i^*, T_i)$ to

$$TII(i) = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} d(T_i^{*(k)}, T_i^{(l)}). \tag{6.1}$$

There are several unresolved issues with this way of doing things. First and as discussed in Chapter D for Bayesian inference, with this definition, TII is not a distance anymore. Second, there is no clear cut answer about how to select the set of trees. It would of course be nice if the set was the confidence region of the ML tree. Several methods are available in the literature to build confidence regions from a set of candidate trees: the SH test (Shimodaira and Hasegawa, 1999), the AU test (Shimodaira, 2002), the LW test (Strimmer and Rambaut, 2002) and all those mentioned in Section 2.4.3. All of them roughly behave the same way; starting from the ML tree, they gradually expand the region by adding trees not significantly different from the ML tree, until a given coverage is reached. But there is no guidelines as to how to choose the initial candidates. If the set of initial candidates is too big, it will be hard for SH to reject the tested trees. Likewise, if the set of candidate trees is not chosen adequately, LW may end up adding all trees to the confidence region. Third, it is not obvious that giving equal weights to all trees in the confidence set is the best choice, posterior probabilities or BP values could be used instead. Finally, if $(T_i^{*(k)})_{k=1..K}$ and $(T_i^{(l)})_{l=1..L}$ span too large a region of the tree space, the influence of species i on the inference will be swallowed up in their span.

Sites × species TII and IF values provide information about variability of the estimates at the sites or species level, by pruning either lines or columns from the alignment. But we can keep going to a even finer scale and prune elements of the matrix to see the influence of a specific nucleotide (or amino-acid but for the sake of simplicity, we use only nucleotides here) on the estimates. Unlike species or sites, deleting a single element would leave a hole in the matrix, but holes have a specific meaning in the alignment (coded by '-') and correspond to gaps, treated as missing data. The modification must therefore be carried out differently.

The first option is to use the empirical influence function (EIF) of the log-likelihood statistic. This means replacing the nucleotide of interest by some arbitrarily value and looking at the output of the estimator. Formally, we define:

$$\begin{array}{ccccc} & \mathcal{M}_{s,n}(\{A,C,G,T\}) & & \to & \mathcal{M}_{s,n}(\{A,C,G,T\}) \\ U^{ij,x}: & \mathbf{X} = \begin{pmatrix} X_{11} & \dots & X_{1j} & \dots & X_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & \dots & X_{ij} & \dots & X_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{s1} & \dots & X_{sj} & \dots & X_{sn} \end{pmatrix} & \mapsto & U^{ij,x}(\mathbf{X}) = \begin{pmatrix} X_{11} & \dots & X_{1j} & \dots & X_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{i1} & \dots & x & \dots & X_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ X_{s1} & \dots & X_{sj} & \dots & X_{sn} \end{pmatrix} \end{array}$$

where $x \in \{A, C, G, T\}$, $i \leq s$ and $j \leq n$. $U^{ij,x}$ merely replaces the (i, j)-th coordinate of a matrix by x. The empirical influence function is then

$$EIF_{ij}(x) = S(U^{ij,x}(\mathbf{X})). \tag{6.2}$$

The set $\{EIF_{ij}(A), \dots, EIF_{ij}(T)\}$ gives us all possible values of $EIF_{ij}(x)$. High variance in the set implies high sensibility to nucleotide X_{ij} whereas no variations means no difference, in terms of log-likelihood, between nucleotides.

However this not yet enough to qualify an element as a outlier. Imagine $EIF_{ij}(A)$ is very high and all other values very low. If we expect with high probability to find an A in this position, observing an A does not come as a surprise. If we observe a T instead, we can be surprised. We must thus assess the expected value of EIF_{ij}. Like in Cook's distance (Cook, 1979) in linear regression, a large value $EIF_{ij}(X_{ij})$ is not informative in itself and needs to be compared to a meaningful value, $E[EIF_{ij}(X)]$ here.

Assuming we do not have any information about X_{ij} and drawing it under the stationary distribution of the evolution model to compute $E[EIF_{ij}(X)]$ would not take into account the tree structure hidden in the alignment and the correlations between $X_{1j}, \ldots, X_{sj}$ and X_{ij}. If all coordinates of the j-th column are A, X_{ij} is also likely to be A (see Fig. 6.2). Given the rest of the alignment, the model M and the ML tree $\hat{T}$, the probabilities $p_x = P(X_{ij} = x)$ are easy to compute in the following way:

$$p_x = \frac{P\left(U^{ij,x}(\mathbf{X}); \hat{T}, M\right)}{\sum_{y \in \{A,C,G,T\}} P\left(U^{ij,y}(\mathbf{X}); \hat{T}, M\right)}$$

and $E[EIF_{ij}(X)]$ should be compared to

$$E[EIF_{ij}(X_{ij})] = \sum_{x \in \{A,C,G,T\}} p_x EIF_{ij}(x).$$

Note that the p_x are computed from the same tree (the ML tree $\hat{T}$) while a new ML tree must be inferred for the computation of each $EIF_{ij}(x)$.

There are at least two potential drawbacks to the nucleotide level approach. First, they are ns elements in the alignment, compared to s species and n nucleotides. Computing $EIF_{ij}(x)$ is too time-consuming to perform an exhaustive search and we must thus use some guidelines as to the potentially interesting nucleotides. Nucleotides belonging to influent species and/or sites are potential candidates. The implicit assumption here is of course that the nucleotide is so influent, it makes both (or either) its sites and species influent. The second concern is lack of power. Substituting a nucleotide to another is a very light perturbation of the alignment and might not tell us a lot about the nucleotide under consideration. It might be sensible to delete (rectangular) blocks of nucleotides instead of single ones. Block pruning mimics the effect of incomplete sampling when the same portion of the alignment (for example a gene) is missing in several species.

Mixture models and artefactual signal Another potential concern with IF and TII values is the presence of artefactual signal. We detect outliers through their exceptional IF and TII values. Therefore, exceptional IF and TII values should really be exceptional, not just exceptional because the model is flawed.

Binary variables are again helpful to illustrate the problem. Consider $(X_1, Y_1), \ldots, (X_n, Y_n)$ a series of i.i.d. couple of binary variables. Assume the distribution of (X_1, Y_1) is $P_{p_x,p_y} = \mathcal{B}(p_x) \otimes \mathcal{B}(p_y)$ but modeled by $P_{p,p} = \mathcal{B}(p) \otimes \mathcal{B}(p)$. The ML-estimator of (p, p) is $(\hat{p}, \hat{p}) = (\frac{\bar{X}_n + \bar{Y}_n}{2}, \frac{\bar{X}_n + \bar{Y}_n}{2})$. Suppose for instance $(p_x, p_y) = (0.1, 0.9)$.

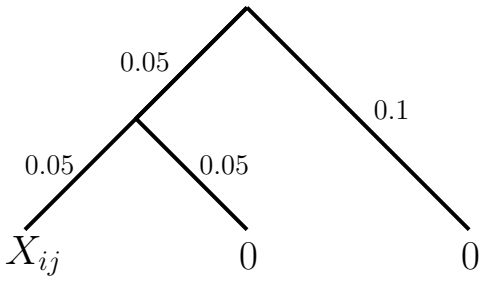

Figure 6.2: The state X_{ij} of the j-th (binary) character of species i is unknown. The stationary distribution of character is $(1/2, 1/2)$ but given the tree and the state of the character for other species, $P(X_{ij} = 0) = 0.89$.

For large n, $\hat{p}$ is close to $1/2$. Deleting X or Y and computing $\hat{p}$ independently on each gives results closer to 0.1 and 0.9 and raises suspicions about X and Y being influent coordinates when the problem lies neither in X nor Y but in the too restrictive model $P_{p,p} = \mathcal{B}(p) \otimes \mathcal{B}(p)$.

This example may seem overreaching; a quick look at the (X_i, Y_i) and very simple diagnostic statistics would prove a model of the form $P_{p,p}$ is not adapted to the problem at hand and should be replaced by a model P_{p_x,p_y}. It is however analogous to the oldest DNA sequence evolution model, the Jukes-Cantor (JC69) (Jukes and Cantor, 1969), which assumes equal mutation rates, not matter which nucleotides bases are involved. Distinction between different transitions and transversions rates came only later (Kimura (1980) and see Sec. 2.2.2).

Artificially exceptional TII and IF values are likely to occur if the real distribution is a mixture distribution. Let us use again a simple example to illustrate the problem. Suppose the (X_i) are i.i.d. variables sampled from a mixture distribution; they take value 0 with probability p and follow a normal distribution $\mathcal{N}(0, \sigma_0^2)$ with probability $1 - p$. Assume the distribution of (X_i) is modeled by $\mathcal{N}(0, \sigma^2)$. The estimator of σ^2 is $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2$. The log-likelihood of F_n is given by:

$$S(F_n) = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\hat{\sigma}_n} e^{-\frac{X_i^2}{2\hat{\sigma}_n^2}}\right) = -\left(\frac{1}{2} + \log\sqrt{2\pi} + \log\hat{\sigma}_n\right).$$

Likewise

$$S(F_n^{(h)}) = -\left(\frac{1}{2} + \log\sqrt{2\pi} + \log\hat{\sigma}_n^{(h)}\right).$$

And finally

$$\begin{aligned}(n-1)\left(S(F_n) - S(F_n^{(h)})\right) &= -(n-1)\log\frac{\hat{\sigma}_n}{\hat{\sigma}_n^{(h)}} \\ &= -\frac{n-1}{2}\log\left(1 + \frac{X_h^2}{\sum_{i\neq h}^n X_i^2}\right) + (n-1)\log\frac{n}{n-1} \xrightarrow[n\to\infty]{\mathcal{L}} 1 - \frac{X_h^2}{2(1-p)\sigma_0^2}\end{aligned}$$

If X_h is sampled from the Dirach mass at 0, $IF_{S,F_n}(X_h) \simeq 1$. If X_h is sampled from the normal distribution, $2(1-p)(1 - IF_{S,F_n}(X_h))$ follows a χ_1^2 distribution. If we declare the $100\alpha\%$ most extreme values as influent values, they correspond under the mixture

model to $|X_h| \geq F^{-1}\left(1 - \frac{\alpha}{2(1-p)}\right)\sigma_0$, or $IF(X_h) \leq 1 - \frac{F^{-1}\left(1-\frac{\alpha}{2(1-p)}\right)^2}{2(1-p)}$, where F is the inverse cumulative distribution function of $\mathcal{N}(0,1)$. Note that $\alpha \leq (1-p)$ — otherwise, all non-null observations would be among the $100\alpha\%$ most extremes. In the absence of mixture ($p = 0$), it simplifies to $IF(X_h) \leq 1 - \frac{F^{-1}\left(1-\frac{\alpha}{2}\right)^2}{2}$. Using the no-mixture IF threshold leads to labeling observations as influent as soon as $|X_h| \geq (1-p)^{1/2}F^{-1}\left(1 - \frac{\alpha}{2}\right)\sigma_0$. For $p = 0.3$ and $100\alpha = 5\%$, 10% of the observations are labeled influent, twice more than the calibration. In conclusion, if IF threshold is set up with no mixture, the mixture induces points drawn from $\mathcal{N}(0, \sigma_0^2)$ are labeled as influent too often. Note that the higher p, the more impact each point drawn from $\mathcal{N}(0, \sigma_0^2)$ has and the lower $IF(X_h)$ should be to qualify X_h as influent. Intuitively, considering in this example that all X_i are drawn from the same distribution artificially reduces the typical heterogeneity of X_i and makes X_i drawn from the more variable distribution look influent.

This example is again too artificial and simple to be realistic but mimics heterogeneity of rates across sites (RAS) (Yang (1994a) and see paragraph 2.2.3). Sites along the sequence do no all evolve at the same speed, some are fast and some are invariant, just like the components of the mixture had positive and null variance. RAS is a concern if influent sites just happen to be fast-evolving sites. We must therefore at least assign a mutation rate to each site of the alignment, using for instance a gamma distribution of rates across sites (+Γ, see Section 2.2.3). We must then test the correlation between mutation rate and IF value. If correlation turns out to be true, we must change the model to incorporate rate heterogeneity.

An alternative, suggested by the segmented curve of IF values (Fig. C.1)in the application of Chapter C, could be to use IF values to segment the sequence and use different models on different portions of the sequence. This is particularly useful for cluster of sites with high IF values, as such a pattern hints that the model used to analyze the sequence may not be relevant for the cluster.

General Conclusion

Chapter 7

Conclusion

7.1 Summary

We focused in this thesis on several aspects of phylogenetic inference. We focused more specifically on validating the estimated tree by quantifying the variability of the estimate. Variability arises for a number of reasons. We developed methods to pinpoint and quantify the contribution of a few of them. We divided the sources of variability into two groups: variability induced by the stochastic process and variability induced by outliers. We studied the former in Part I and the latter in Part II.

- The first source of variability, studied in Part I, is induced by the sampling of observations and is of course inherent to any statistical inference process. The likelihood of a tree is an essential quantity since it is used to rank trees from best to worst and to select the maximum likelihood tree. It is calculated from the columns of the alignment matrix. The columns are i.i.d random variables drawn from some distribution. Hence, the likelihood scores is itself a random variable. But as the number of columns increases, it converges to its asymptotic value. We studied in Chapter A how fast the likelihood-score of a tree converges to its asymptotic value and gave Chernoff-like results.
- The second source of variability, studied in Part I, is induced by the lack of consistency of the site distribution along the sequence. This can be thought as a change of the evolutionary process, and hence of the phylogenetic signal along the sequence. If the distribution change comes with a shift in the expected log-likelihood per site, this introduces a bias in the estimated likelihood score of a tree. We studied in Chapter B how shifts in the expected likelihood per site can be detected. We proposed a complete procedure, based on Edgeworth expansions and resampling techniques, to detect and test significant shifts.
- The third source of variability, studied in Part II, is induced by influent sites whose inclusion/exclusion from the analysis strongly affects the estimates. Influent sites can be genuinely influent and contribute a more than welcome strong phylogenetic signal to the inference. They can also be outliers resulting from sequencing/alignment errors and hinder the inference. We studied detection of influent sites in Chapter C. We proposed a index, inspired by influence function and sensitivity curve, to quantify the influence of a site and highlight potential

outliers.
– The fourth source of variability, studied in Part II, is induced by influent species whose inclusion/exclusion from the analysis strongly affects the estimates. Influent species can be genuinely influent and have a stabilizing effect on the phylogeny. They can also be rogue taxa which are hard to resolve and hinder the inference. We studied detection of influent species in Chapter D. It is similar to detecting influent sites but has less statistical founding. We proposed simple index, the Taxon Influence Index, based on tree distances to quantify the influence of a species and highlight potential rogue taxa.

7.2 Perspectives

Phylogenetics in general and the robustness of phylogenetic trees in particular is a very broad field which calls for more attention and research than this thesis could possibly cover. Generalizations of our results or possible subjects for further research are described and discussed in Chapters 4 and 6 and are not fleshed out here again. Some call for immediate attention while some are long term prospects. We focus here only on the work we want to do in a reasonable future.

First and foremost, we need to broadcast the IF and TII indexes to a broader audience. The indexes are easy to understand and once understood little work is needed to calculate them but inclusion to a phylogenetic analysis packaqe would make it more appealing to the end-user. The most popular R package for phylogenetic inference manipulation and analyses are `ape` (Paradis et al., 2004) and `phangorn`, which makes them obvious candidates.

In line with this idea, we need to test our indexes on more data sets than just the two used in Chapter C and Chapter D. This would allow to gain knowledge about the general behavior of IF and TII values. Data sets suspected to exhibit strong sensitivity to taxon and sites sampling are good candidates. One such example is Rokas et al. (2003) yeast dataset. Another promising example would be reptilians (Jonniaux and Kumazawa (2008) and private communication with P. Jonniaux).

Combining IF and TII to look for influent sites of influent species is the logical continuation. It would allow us to pinpoint sources of variability at an even finer scale than sites or species. This strategy is outlined and discussed in the paragraph "sites × species" of section 6.2.

IF are inspired by influence function while TII are based on simple similarity measures. For now, the threshold for labeling a site (or a taxon) as influent is set a bit arbitrarily. Setting the threshold on more solid theoretical grounds or finding a credible way to set the threshold is desirable. This issue is discussed in the paragraph "setting a threshold" of section 6.2.

The next concern about IF and TII values is their assumption that the inference method is indeed able to find the ML tree. This is not obvious as we saw in section 2.3. This caveat is of course not specific to IF and TII values but shared by many robustness indexes, most notably bootstrap values. If the ML tree is not significantly better than the second best, it could be interesting to consider both instead of just the ML tree

and use similarity measures between sets of trees rather than single trees. This is discussed in more details in the paragraph "setting a threshold" of section 6.2.

Studying sets of trees instead of single trees immediately raise the issue of adequate distance between trees and their adaptations to set of trees. Although several distances between sets of points are used daily in classification, there is to our knowledge no distance between sets of trees that takes the tree structure into account and quantifies how different the sets are in a satisfying way. This is however an absolute necessity to start working with sets of trees without aggregating them in consensus tree or a phylogenetic network.

On a different note, independence of sites is a well-worn issue in phylogenetics. There have been tries at inferring distances (Falconnet, 2008) or evolutionary parameters (Arndt et al., 2003; Siepel and Haussler, 2004) from non independent sites but this is not our concern. The general feeling in the field is that dependence reduces the "effective" number of sites. Bootstrap and other resampling methods assumes that the sites are independent and may be too confident when they are not. The most unrealistic yet simplest illustration is n identical copies of the same observation. Resampling in the original data set always gives the same data set and we are thus overconfident that the estimate is correct. It could be interesting to estimate the "effective" sample size and resample only that number of sites.

Finally, inferring phylogenies and validating them is worthy of interest per se but phylogenies can also be (and actually are) interesting for other goals. Functional divergence is an interesting example. Functional divergence occurs after a gene duplication event. Type I divergence results in altered functional constraints while type II divergence results not in altered functional constraints but in changes in amino-acids properties. Phylogenies allow us to study this evolution (Gu, 2001, 2006) but statistical methods inspired from spatial statistics such as quadrant analysis could also be adapted to the problem.

Bibliography

F. Abascal, R. Zardoya, and D. Posada. Prottest: selection of best-fit models of protein evolution. *Bioinformatics*, 21(9):2104–2105, May 2005.

B. L. Allen and M. Steel. Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, 5(1):1–15, 2001.

J. M. Archibald and A. J. Roger. Gene conversion and the evolution of euryarchaeal chaperonins: a maximum likelihood-based method for detecting conflicting phylogenetic signal. *J Mol Evol*, 55:232–245, 2002.

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significance tests on accumulating data. *J. Roy. Statist. Soc. Ser. A*, 132(2):235–244, 1969.

P. F. Arndt, C. B. Burge, and T. Hwa. Dna sequence evolution with neighbor-dependent mutation. *J Comput Biol*, 10(3-4):313–322, 2003.

D. J. Balding, M. Bishop, and C. Cannings, editors. *Handbook of Statistical Genetics, 3rd Edition*. Wiley, 2007.

A. Bar-Hen and H. Kishino. Comparing the likelihood functions of phylogenetic trees. *Annals of the Institute of Statistical Mathematics*, 52(1):43–56, March 2000.

A. Bar-Hen, M. Mariadassou, M.-A. Poursat, and P. Vandenkoornhuyse. Influence function for robust phylogenetic reconstructions. *Mol Biol Evol*, 25(5):869–873, May 2008.

N. M. Belfiore, L. Liu, and C. Moritz. Multilocus phylogenetics of a rapid radiation in the genus thomomys (rodentia: geomyidae). *Syst Biol*, 57(2):294–310, Apr 2008.

G. Bello, C. Casado, S. García, C. Rodríguez, J. del Romero, A. Carvajal-Rodriguez, D. Posada, and C. López-Galíndez. Lack of temporal structure in the short term hiv-1 evolution within asymptomatic naïve patients. *Virology*, 362(2):294–303, Jun 2007.

G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962. ISSN 01621459.

A. Benveniste, M. Basseville, and G. Moustakides. The asymptotic local approach to change detection and model validation. *IEEE Trans. Automatic Control*, 32:583–592, 1987.

L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Math.*, 27:733–767, 2001.

S. Blanquart and N. Lartillot. A bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol Biol Evol*, 23(11): 2058–2071, Nov 2006.

S. Blanquart and N. Lartillot. A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol*, 25(5):842–858, May 2008.

C. Blouin, D. Butt, and A. J. Roger. The impact of taxon sampling on the estimation rate of evolution at sites. *Mol Biol Evol*, 22:784–791, 2005.

M. Bordewich, A. G. Rodrigo, and C. Semple. Selecting taxa to save or sequence: desirable criteria and a greedy solution. *Syst Biol*, 57(6):825–834, Dec 2008.

B. Boussau and M. Gouy. Efficient likelihood computations with nonreversible models of evolution. *Syst Biol*, 55(5):756–768, Oct 2006.

B. Brodsky and B. Darkovsky. *Non-Parametric Statistical Diagnosis. Problems and Methods*. Kluwer, Dordrecht, 2000.

B. Brodsky and B. Darkovsky. Asymptotically optimal methods of change-point detection for composite hypotheses. *Journal of Statistical Planning and Inference*, 133: 123–138, 2005.

D. Bryant, N. Galtier, and M. A. Poursat. *Mathematics of evolution and phylogeny*, chapter Likelihood calculation in molecular phylogenetics, pages 33–62. Oxford University Press, 2005.

T. R. Buckley. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst Biol*, 51(3):509–523, Jun 2002.

J. Buschbom and A. von Haeseler. *Statistical Methods in Molecular Evolution*, chapter Introduction to Applications of the Likelihood Function in Molecular Evolution, pages 25–44. Springer, 2004.

R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch. Predicting the evolution of human influenza a. *Science*, 286(5446):1921–1925, Dec 1999.

J. Bérard, J.-B. Gouéré, and D. Piau. Solvable models of neighbor-dependent substitution processes. *Math Biosci*, 211(1):56–88, Jan 2008.

Y. Cao, J. Adachi, T. Yano, and M. Hasegawa. Phylogenetic place of guinea pigs: no support of the rodent-polyphyly hypothesis from maximum-likelihood analyses of multiple protein sequences. *Mol Biol Evol*, 11(4):593–604, Jul 1994.

Y. Cao, N. Okada, and M. Hasegawa. Phylogenetic position of guinea pigs revisited. *Mol Biol Evol*, 14(4):461–464, Apr 1997.

L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetics analysis: Models and estimation procedures. *American Journal of man Geneics*, 19:233–257, 1967.

A. Chao. Nonparametric estimation of the number of classes in a population. *Scand. J. Statis.*, 11:265–270, 1984.

R. D. Cook. Influential observations in linear regression. *J. Amer. Statist. Assoc.*, 74 (365):169–174, 1979. ISSN 0003-1291.

H. Cramer. *Random Variables and Probability Distributions*, volume 36 of *Cambrige Tracts in Mathematics*. Cambridge University Press, 1937.

K. A. Cranston and B. Rannala. Summarizing a posterior distribution of trees using agreement subtrees. *Syst Biol*, 56(4):578–590, Aug 2007.

A. Cuevas and J. Romo. On the estimation of the influence curve. *The Canadian Journal of Statistics*, 23(1):1–9, 1995.

C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (1st ed.), London: John Murray,*. John Murray, London, 1 edition, 1859.

A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications, Second Edition*, volume 38 of *Application of Mathematics*. Springer, 2 edition, March 1998.

A. M. D'Erchia, C. Gissi, G. Pesole, C. Saccone, and U. Arnason. The guinea-pig is not a rodent. *Nature*, 381(6583):597–600, Jun 1996.

P. Domingos and G. Hulten. Mining high-speed data streams. *KDD*, pages 71–80, 2000.

C. W. Dunn, A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sørensen, S. H. D. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q. Martindale, and G. Giribet. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452(7188):745–749, Apr 2008.

L. Duret. The gc content of primates and rodents genomes is not at equilibrium: a reply to antezana. *J Mol Evol*, 62(6):803–806, Jun 2006.

L. Duret. Neutral theory: The nulll hypothesis of molecular evolution. *Nature Education*, 1:1, 2008.

L. Duret and N. Galtier. The covariation between tpa deficiency, cpg deficiency, and g+c content of human isochores is due to a mathematical artifact. *Mol Biol Evol*, 17 (11):1620–1625, Nov 2000.

A. W. F. Edwards. Estimation of the branch points of a branching diffiasion process (with discussion). *J. R. Stat. Soc. B*, 32:155–174, 1970.

A. W. F. Edwards and L. L. Cavalli-Sforza. The reconstruction of evolution. *Annals of Human Genetics*, 27:105–106, 1963.

A. W. F. Edwards and L. L. Cavalli-Sforza. *Phenetic and Phylogenetic Classification*, chapter Reconstruction of evolutionary trees, pages 67–76. Systematics Association Publ. No. 6, London, 1964.

B. Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1): 1–26, January 1979. 2.0.CO

B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci U S A*, 93(14):7085–7090, Jul 1996.

J. A. Eisen. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res*, 8(3):163–167, Mar 1998.

J. A. Eisen and M. Wu. Phylogenetic analysis and gene functional predictions: phylogenomics in action. *Theor Popul Biol*, 61(4):481–487, Jun 2002.

C.-G. Esseen. Fourier analysis of distribution functions. A mathematical study of the Laplace-Gaussian law. *Acta Math.*, 77:1–125, 1945. ISSN 0001-5962.

M. Falconnet. Phylogenetic distances for neighbour dependent substitution processes. 2008. `http://arxiv.org/abs/0812.1962`

J. Farris. *Advances in Cladistics*, chapter The Logical Basis of Phylogenetic Analysis., pages 7–36. Columbia University Press, New York, 1983.

J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, September 2004.

J. Felsenstein. *Statistical Inference and the Estimation of Phylogenies*. PhD thesis, Department of Zoology, University of Chicago, 1968.

J. Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 22:240–249, 1973.

J. Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410, 1978.

J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

J. Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, 39(4):783–791, July 1985.

J. Felsenstein. Phylip (phylogeny inference package) version 3.6. Distributed by the author, 2005.

J. Felsenstein and H. Kishino. Is there something wrong with the bootstrap on phylogenies? a reply to hillis and bull. *Systematic Biology*, 42(2):193–200, June 1993.

J. Felsentstein. Statistical inference of phylogenies. *J.R. Statist. Soc.*, 146(3):246–272, 1983.

R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.

W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155(760): 279–284, Jan 1967.

N. Galtier. Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol*, 18(5):866–873, May 2001.

N. Galtier and L. Duret. Adaptation or biased gene conversion? extending the null hypothesis of molecular evolution. *Trends Genet*, 23(6):273–277, Jun 2007.

O. Gascuel. *Mathematics of evolution and phylogeny*. Oxford University Press, 2005.

J. Gatesy, R. DeSalle, and N. Wahlberg. How many genes should a systematist sample? conflicting insights from a phylogenomic matrix characterized by replicated incongruence. *Syst Biol*, 56(2):355–363, Apr 2007.

K. Geuten, T. Massingham, P. Darius, E. Smets, and N. Goldman. Experimental design criteria in phylogenetics: where to add taxa. *Syst Biol*, 56(4):609–622, Aug 2007.

A. L. Gibbs and E. Su. On choosing and bounding probability metrics. *Intl. Stat. Rev.*, 7(3):419–435, 2002.

N. Goldman. Simple diagnostic statistical tests of models for dna substitution. *J Mol Evol*, 37(6):650–661, Dec 1993a.

N. Goldman. Statistical tests of models of dna substitution. *J Mol Evol*, 36(2):182–198, Feb 1993b.

N. Goldman. Phylogenetic information and experimental design in molecular systematics. *Proc Biol Sci*, 265(1407):1779–1786, Sep 1998.

N. Goldman, J. P. Anderson, and A. G. Rodrigo. Likelihood-based tests of topologies in phylogenetics. *Syst. Biol.*, 49(4):652–670, Dec 2000.

D. Graur, W. A. Hide, and W. H. Li. Is the guinea-pig a rodent? *Nature*, 351(6328): 649–652, Jun 1991.

I. Gronau, S. Moran, and S. Snir. Fast and Reliable Reconstruction of Phylogenetic Trees with Very Short Edges. In *SODA: ACM-SIAM Symposium on Discrete Algorithms*, pages 379–388, 2008.

X. Gu. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol*, 18(4):453–464, Apr 2001.

X. Gu. Statistical framework for phylogenomic analysis of gene family expression profiles. *Genetics*, 167(1):531–542, May 2004.

X. Gu. A simple statistical method for estimating type-ii (cluster-specific) functional divergence of protein sequences. *Mol Biol Evol*, 23(10):1937–1945, Oct 2006.

X. Gu, Z. Zhang, and W. Huang. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci U S A*, 102(3):707–712, Jan 2005.

S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696–704, Oct 2003.

H. Guo, R. E. Weiss, X. Gu, and M. A. Suchard. Time squared: repeated measures on phylogenies. *Mol Biol Evol*, 24(2):352–362, Feb 2007.

P. Hall. *Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer, 1984.

F. R. Hampel. The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69:383–393, 1974a.

F. R. Hampel. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.*, 69:383–393, 1974b.

M. Hasegawa and M. Fujiwara. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol Phylogenet Evol*, 2(1):1–5, Mar 1993.

M. Hasegawa and H. Kishino. Confidence limits on the maximum likelihood estimate of the hominoid tree from mitochondrial dna sequences. *Evolution*, 43:672–677, 1989.

M. Hasegawa and H. Kishino. Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree. *Mol Biol Evol*, 11:142–145, 1994.

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial-dna. *Journal of Mol Biol Evol*, 22:160–174, 1985.

T. A. Heath, S. M. Hedtke, and D. M. Hillis. Taxon sampling and the accuracy of phylogenetic analyses. *J Mol Evol*, 46:239–257, 2008.

P. D. N. Hebert, A. Cywinska, S. L. Ball, and J. R. deWaard. Biological identifications through dna barcodes. *Proc Biol Sci*, 270(1512):313–321, Feb 2003.

S. M. Hedtke, T. M. Townsend, and D. M. Hillis. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst Biol*, 55(3):522–529, Jun 2006.

D. M. Hillis and J. J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, 42(2):182–192, June 1993.

D. M. Hillis, D. D. Pollock, J. A. McGuire, and D. J. Zwickl. Is sparse taxon sampling a problem for phylogenetic inference? *Syst Biol*, 52(1):124–126, Feb 2003.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. ISSN 01621459.

B. R. Holland, D. Penny, and M. D. Hendy. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock–a simulation study. *Syst Biol*, 52(2):229–238, Apr 2003.

S. Holmes. *Mathematics of evolution and phylogeny*, chapter Statistical approach to tests involving phylogenies, pages 91–120. Oxford University Press, 2005.

S. Holmes. Bootstrapping phylogenetic trees: Theory and methods. *Statistical Science*, 18:241–255, 2003.

P. J. Huber. *Robust Statistics*. Wiley, New York, 1981.

P. J. Huber. *Robust Statistics*. Wiley, Sussex, 2004.

J. P. Huelsenbeck. Testing a covariotide model of dna substitution. *Mol Biol Evol*, 19 (5):698–707, May 2002.

J. P. Huelsenbeck and J. P. Bollback. *Handbook of Statistical Genetics, 3rd Edition*, chapter Application of the Likelihood Function in Phylogenetic Analysis, pages 460–488. Wiley, 3 edition, 2007.

A. J. Jeffreys, V. Wilson, and S. L. Thein. Hypervariable 'minisatellite' regions in human dna. *Nature*, 314(6006):67–73, 1985.

D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8(3):275–282, Jun 1992.

P. Jonniaux and Y. Kumazawa. Molecular phylogenetic and dating analyses using mitochondrial dna sequences of eyelid geckos (squamata: Eublepharidae). *Gene*, 407(1-2):105–115, Jan 2008.

T. Jukes and C. Cantor. *Evolution of protein molecules*, volume 3 of *mammalian Protein Metabolism*, chapter 24, pages 21–132. Academic Press, New York, 1969.

M. G. Kendall and A. Stuart. *The advanced theory of statistics*. Charles Griffin, London, 1973.

J. Kim. Large-scale phylogenies and measuring the performance of phylogenetic estimators. *Syst Biol*, 47(1):43–60, Mar 1998.

J. Kim. Slicing hyperdimensional oranges: the geometry of phylogenetic estimation. *Mol Phylogenet Evol*, 17(1):58–75, Oct 2000.

M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, Feb 1968.

M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, Dec 1980.

H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *J Mol Evol*, 29(2):170–179, Aug 1989.

Y. Kitazoe, H. Kishino, P. J. Waddell, N. Nakajima, T. Okabayashi, T. Watabe, and Y. Okuhara. Robust time estimation reconciles views of the antiquity of placental mammals. *PLoS ONE*, 2(4):e384, 2007.

M. K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol*, 11(3):459–468, May 1994.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, Mar 1951.

H. R. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241, 1989.

S. N. Lahiri. Theoretical comparisons of block bootstrap methods. *The Annals of Statistics*, 27(1):386–404, 1999.

T. L. Lai. Sequential changepoint detection in quality control and dynamical systems (with discussion). *J. Roy. Statist. Soc. Ser. B.*, 57:613–658, 1995.

T. L. Lai and J. Z. Shan. Efficient recursive algorithms for detection of abrupt changes in signals and systems. *IEEE Trans. Automatic Control*, 44:952–966, 1999.

C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *J Mol Evol*, 20(1):86–93, 1984.

S. M. Lanyon. Detecting internal inconsistencies in distance data. *Syst Zool*, 34(4): 397–403, Dec 1985.

Larson. *Evolution: The Remarkable History of a Scientific Theory*. Modern Library, 2006.

N. Lartillot and H. Philippe. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol*, 21(6):1095–1109, Jun 2004.

N. Lartillot, H. Brinkmann, and H. Philippe. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*, 7 Suppl 1:S4, 2007.

G. Lecointre, H. Philippe, H. L. V. Lê, and H. L. Guyader. Species sampling has a major impact on phylogenetic inference. *Mol Phylogenet Evol*, 2(3):205–224, Sep 1993.

C. Li, G. Lu, and G. Orti. Optimal data partitioning and a test case for ray-finned fishes (actinopterygii) based on ten nuclear loci. *Syst Biol*, 57(4):519–539, Aug 2008.

S. Li. *Phylogenetic tree construction using Markov chain Monte Carlo*. PhD thesis, Ohio State University, Ohio, 1996.

K. Liu, S. Raghavan, S. Nelesen, C. R. Linder, and T. Warnow. Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, 324(5934):1561–1564, Jun 2009.

C. X. Mao and B. G. Lindsay. Estimating the number of classes. *The Annals of Statistics*, 35(2):917–930, 2007.

P. Massart. *Concentration inequalities and model selection*. Springer, 2006.

F. A. Matsen and M. Steel. Phylogenetic mixtures on q single tree can mimic a tree of another topology, 2007. `http://www.citebase.org/abstract?id=oai:arXiv.org:0704.2260`.

F. A. Matsen, E. Mossel, and M. Steel. Mixed-up trees: the structure of phylogenetic mixtures, 2007. `http://www.citebase.org/abstract?id=oai:arXiv.org:0705.4328`.

B. Mau. *Bayesian phylogenetic inference via Markov chain Monte Carlo*. PhD thesis, University of Wisconsin, 1996.

G. J. Mendel. Versuche über plflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865*, Abhandlungen:3–47, 1866.

C. D. Michener and R. R. Sokal. A quantitative approach to a problem in classification. *Evolution*, 11:130–162, 1957.

R. G. Miller. The jackknife — A review. *Biometrika*, 61(1):1–15, 1974.

E. Mossel and M. Steel. *Mathematics Of Evolution And Phylogeny*, chapter How much can evolved characters tell us about the tree that generated them?, pages 384–312. Oxford University Press, 2005.

M. Newton. Bootstrapping phylogenies: Larges deviations and dispersion effects. *Biometrika*, 83(2):315–358, 1996.

R. Nielsen. *Statistical Methods in Molecular Evolution*. Statistics for Biology and Health. Springer, 2004.

E. Nowak and A. Bar-Hen. Influence function and correspondence analysis. *Journal of Statistical Planning and Inference*, 134(1):26–35, 2005.

P. C. O'Brien and T. R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556, 1979.

R. Ota, P. J. Waddell, M. Hasegawa, H. Shimodaira, and H. Kishino. Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Mol Biol Evol*, 17(5):798–803, May 2000.

S. P. Otto, M. P. Cummings, and J. Wakeley. *New Uses for New Phylogenies*, chapter Inferring phylogenies from DNA sequence data: the effects of sampling. Oxford University Press, 1996.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41:100–114, 1954.

B. Papp, C. Pál, and L. D. Hurst. Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet*, 19(8):417–422, Aug 2003.

E. Paradis, J. Claude, and K. Strimmer. Ape: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290, 2004.

F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9):609–614, Sep 2001.

J. Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *AAAI-82 Proceedings*, Pittsburgh, PA, 1982. AAAI Press.

D. Penny and M. Hendy. Estimating the reliability of evolutionary trees. *Molecular Biology and Evolution*, 3:403–417, 1986.

D. Penny and M. D. Hendy. The use of tree comparison metrics. *Systematic Zoology*, 34:75–82, 1985. TCM

D. Penny, M. Hendy, and M. Steel. Progress with evolutionary trees. *Trends Ecol Evol*, 7:73–79, 1992.

H. Philippe. Rodent monophyly: pitfalls of molecular phylogenies. *J Mol Evol*, 45(6): 712–715, Dec 1997.

S. J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–199, 1977.

S. Poe. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. *Syst Biol*, 52(3):423–428, Jun 2003.

S. Poe and D. L. Swofford. Taxon sampling revisited. *Nature*, 398(6725):299–300, Mar 1999.

D. D. Pollock, D. J. Zwickl, J. A. McGuire, and D. M. Hillis. Increased taxon sampling is advantageous for phylogenetic inference. *Syst Biol*, 51(4):664–671, Aug 2002.

D. Posada and T. R. Buckley. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst Biol*, 53(5):793–808, Oct 2004.

D. Posada and K. A. Crandall. Modeltest: testing the model of dna substitution. *Bioinformatics*, 14:817–818, 1998.

A. B. Prasad, M. W. Allard, N. I. S. C. C. S. Program, and E. D. Green. Confirming the phylogeny of mammals by use of large comparative sequence datasets. *Mol Biol Evol*, 25:1795–1808, Sep 2008.

B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol*, 43(3):304–311, Sep 1996.

B. Rannala, J. P. Huelsenbeck, Z. Yang, and R. Nielsen. Taxon sampling and the accuracy of large phylogenies. *Syst Biol*, 47(4):702–710, Dec 1998.

D. F. Robinson and L. R. Foulds. *Lectures Note in mathematics*, volume 748, chapter Comparison of weighted labelled trees, pages 119–126. Springer-Verlag, Berlin, 1979.

A. Rokas and S. B. Carroll. More genes or more taxa? the relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol*, 22(5):1337–1344, May 2005.

A. Rokas, B. L. Williams, N. King, and S. B. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, Oct 2003.

F. Ronquist and J. P. Huelsenbeck. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, Aug 2003.

M. S. Rosenberg and S. Kumar. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc Natl Acad Sci U S A*, 98(19):10751–10756, Sep 2001.

T. Sato, Y. Yamanishi, K. Horimoto, M. Kanehisa, and H. Toh. Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions. *Bioinformatics*, 22(20):2488–2492, Oct 2006.

S. G. Self and K.-Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987. ISSN 01621459.

L. Shavit, D. Penny, M. D. Hendy, and B. R. Holland. The problem of rooting rapid radiations. *Mol Biol Evol*, 24(11):2400–2411, Nov 2007.

H. Shimodaira. An approximately unbiased test of phylogenetic tree selection. *Syst Biol*, 51(3):492–508, Jun 2002.

H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol*, 16(8):1114–1116, 1999.

A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol*, 21(3):468–488, Mar 2004.

A. Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, Nov 2006.

M. A. Steel and L. A. Szekely. Inverting random functions. *Annals of Combinatorics*, 3: 103–113, 1999.

M. A. Steel and L. A. Szekely. Inverting random functions (ii): explicit bounds for discrete maximum likelihood estimation, with applications. *SIAM J. Discr. Math.*, 15(4):562–575, 2002.

M. A. Steel and L. A. Szekely. Inverting random functions iii: Discrete mle revisited. August 2006a.

M. A. Steel and L. A. Szekely. On the variational distance of two trees. *The Annals of Applied Probability*, 16(3):1563–1575, 2006b.

K. Strimmer and A. Rambaut. Inferring confidence sets of possibly misspecified gene trees. *Proc Biol Sci*, 269(1487):137–142, Jan 2002.

Study. The beta-blocker heart attack trial. beta-blocker heart attack study group. *JAMA*, 246(18):2073–2074, Nov 1981.

J. Sullivan and D. L. Swofford. Are guinea pigs rodents? the importance of adequate models in molecular phylogenetics. *Journal of Mammalian Evolution*, 4(2):77–96, June 1997.

E. Susko. Bootstrap support is not first-order correct. *Syst Biol*, 58, 2009.

D. L. Swofford. *PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4.04beta*. Sinauer Associates, Sunderland, Massachusetts, 2003.

D. L. Swofford, J. L. Olsen, P. J. Waddell, and D. M. Hillis. *Molecular systematics*, chapter Phylogenetic Inference, pages 407–514. Sinauer Associates, 1996.

Tang and Yongqiang. A hoeffding-type inequality for ergodic time series. *Journal of Theoretical Probability*, 20(2):167–176, June 2007. ISSN 0894-9840.

J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, Nov 1994.

Thorley and Wilkinson. Testing the phylogenetic stability of early tetrapods. *J Theor Biol*, 200(3):343–344, Oct 1999.

C. Tuffley and M. Steel. Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences*, 147(1):63–91, January 1997.

S. Van de Geer. *Empirical Process Techniques for Dependent Data*, chapter On Hoeffding's inequality for dependent random variables, pages 161–170. Birkhäuser, Boston, 2002.

P. Vandenkoornhuyse, S. L. Baldauf, C. Leyval, J. Straczek, and J. P. W. Young. Extensive and novel fungal diversity in plant roots. *Science*, 295:2051, 2002.

Wald. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16:117–186, 1945.

J. D. Watson and F. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

M. Wilkinson. Majority-rule reduced consensus trees and their use in bootstrapping. *Mol Biol Evol*, 13(3):437–444, Mar 1996.

S. Wuchty. Evolution and topology in the yeast protein interaction network. *Genome Res*, 14(7):1310–1314, Jul 2004.

Z. Yang. Maximum-likelihood estimation of phylogeny from dna sequences when substitution rates differ over sites. *Mol Biol Evol*, 10(6):1396–1401, Nov 1993.

Z. Yang. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *J Mol Evol*, 39(3):306–314, Sep 1994a.

Z. Yang. Complexity of the simplest phylogenetic estimation problem. *Proceedings of the Royal Society B: Biological Sciences*, 267:109–116, 2000.

Z. Yang. Fair-balance paradox, star-tree paradox, and bayesian phylogenetics. *Mol Biol Evol*, 24(8):1639–1655, Aug 2007.

Z. Yang. Estimating the pattern of nucleotide substitution. *J Mol Evol*, 39(1):105–111, Jul 1994b.

Z. Yang, N. Goldman, and A. Friday. Maximum likelihood trees from dna sequences: a peculiar statistical estimation problem. *Syst. Biol.*, 44(3):384–399, 1995.

Z. Yang, R. Nielsen, and M. Hasegawa. Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol*, 15(12):1600–1611, Dec 1998.

H. Yu, N. M. Luscombe, H. X. Lu, X. Zhu, Y. Xia, J.-D. J. Han, N. Bertin, S. Chung, M. Vidal, and M. Gerstein. Annotation transfer between genomes: protein-protein interologs and protein-dna regulogs. *Genome Res*, 14(6):1107–1118, Jun 2004.

A. Zharkikh and W. H. Li. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. i. four taxa with a molecular clock. *Mol Biol Evol*, 9(6):1119–1147, Nov 1992.

M. Zucker, D. H. Mathews, and D. H. Turner. *RNA Biochemistry and Biotechnology*, chapter Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide, pages 11–43. Kluwer Academic Publishers, 1999.

D. J. Zwickl and D. M. Hillis. Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol*, 51(4):588–598, Aug 2002.