Tree shape in population genetics and phylogeny

Michael G B Blum

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

THESE

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité: "mathématiques appliquées"

préparée au laboratoire de techniques de l'imagerie, de la modélisation et de la cognition (TIMC)
dans le cadre de **l'Ecole Doctorale "mathématiques, sciences et technologies de l'information, informatique"**

présentée et soutenue publiquement

par

Michael BLUM

le 21 octobre 2005

Titre :

De la forme des généalogies en phylogénie et en génétique des populations

---

Directeur de thèse : Olivier FRANCOIS

JURY

M. Bernard PRUM , Président
M. Catherine LAREDO , Rapporteur
M. Robert C. GRIFFITHS , Rapporteur
M. Olivier FRANCOIS , Directeur de thèse
M. Fréderic AUSTERLITZ , Co-encadrant
M. Oscar GAGGIOTTI , Examinateur

Remerciements

Je tiens à remercier le professeur Bernard Prum pour avoir accepté de présider le jury de cette thèse et de s'être déplacé jusqu'à Grenoble en ce jour pluvieux d'octobre.

Je sais gré à Catherine Laredo, directeur de recherche à l'INRA, d'avoir accepté d'être rapporteur et de m'avoir envoyé des commentaires détaillés. Je tiens aussi à remercier Bob Griffiths, professeur à Oxford, non seulement parce qu'il a accepté de porter un jugement sur mon travail mais aussi parce qu'il a eu la délicatesse d'avoir glissé une assiette sous ma prune alors que je m'apprêtais à la poser à même la table et enfreindre du même coup l'étiquette d'Oxford.

Je souhaite témoigner ma gratitude envers Oscar Gagiotti, professeur à l'UJF, pour avoir accepté de faire partie du Jury et devoir ainsi jouer au contorsionniste avec son emploi du temps.

Si j'ai été amené à m'intéresser à la forme des généalogies, c'est grâce à Frederic Austerlitz, chargé de recherche au CNRS. Il m'a accueilli au sein du laboratoire ESE à Orsay, et m'a tout de suite proposé de participer à ses recherches. Je le remercie vivement ainsi qu'Evelyne Heyer, professeur à Paris 7, de m'avoir fait découvrir cette discipline fascinante qu'est la génétique des populations humaines.

Enfin, cette thèse n'aurait sans aucun doute pas la même teneur sans mon directeur de thèse, Olivier François, professeur à l'INPG. Sa curiosité scientifique a toujours été un moteur. Il s'est intéressé aux différentes thématiques que j'ai abordés et s'y est consacré à son tour. Son investissement restera un exemple pour moi.

On imagine bien souvent que la thèse est un âpre travail solitaire, dans mon cas il serait malhonnête de l'affirmer. C'est grâce aux talents conjugués de scientifiques, de collègues, d'amis et de mes parents que je suis parvenu à la finir. Si j'ai voulu effectuer ma thèse dans le domaine des probabilités et des statistiques, c'est en grande partie grâce aux cours dispensés par Jean-Louis Soler, professeur à l'ENSIMAG. Je remercie aussi Svante Janson, professeur à Uppsala, de m'avoir aidé à surmonter certains problèmes mathématiques. Jacques Demongeot, professeur à l'UJF, m'a encouragé, avec toute la passion qu'on lui connaît, à partir dans le Michigan. Je n'oublierai pas tous mes collègues que j'ai abreuvés de paroles : Pierre, Solenn, Benjamin et tous les autres à Orsay ainsi que Loïc, Adrien, Mathieu, Ju-

lien, Andrès à Grenoble. Je les remercie de ne pas m'avoir écouté, d'avoir corrigé mon anglais, mon français... Mes programmes informatiques sont sous copyright Benjamin Sergeant, je le remercie du fond du coeur. Les simulations effectuées par Christophe ainsi que la librairie R développée par Nicolas et Eric ont été des outils précieux, je leur en suis reconnaissant. Ma soutenance de thèse a été liftée par Hervé Guiol, maître de conférence à l'INPG, et Jean-Louis Martiel, chargé de recherche à L'INSERM, qu'ils en soient remerciés. Et comme une thèse dure 3 ans et que la science ça ne nourrit pas (ou peu), j'ai pris durant cette période un nombre incalculable de repas et passé un nombre presque aussi grand de nuits aux frais de la princesse. Je tiens à m'incliner devant toutes les princesses : Simon, les autres colocataires de la rue Thiers, ceux de la rue Moyrand, la famille Cuel, Marie-Anne et Florence.

J'ai été très honoré que ma famille se déplace pour ma soutenance. Je remercie ma grand-mère, mon grand-père, ma tante, mon frère ainsi que mon cousin pour leurs venues.

Je serai bien évidemment toujours débiteur auprès de mes parents. Je n'ai pas toujours su quelle était ma chance d'avoir un père qui avait les pieds sur terre et une mère la tête dans les nuages avec un téléphone collé à l'oreille, son fils au bout du fil.

Table des matières

1 Introduction **9**

1.1 Modèles stochastiques en phylogenèse 11

1.2 Modèles stochastiques en génétique des populations 14

1.3 Présentation des chapitres 17

1.4 English introduction 27

2 On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited **33**

2.1 Introduction 34

2.2 Theory 36

2.2.1 Sackin's statistic 36

2.2.2 Number of cherries 37

2.2.3 The number of subtrees of fixed size 38

2.2.4 New indices 41

2.3 Distribution of indices 41

2.4 Statistical power 43

2.4.1 Biased speciation model 43

2.5 Example 46

3 The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance **51**

3.1 Introduction 52

3.2 The Sackin and the Colless for large phylogenetic trees 54

3.3 Phylogenetic and binary search trees 58

3.4 Yule model 63

3.5 Uniform model 67

4 Fertility inheritance unbalances gene genealogies 75

4.1 Introduction 76
4.2 Tree Balance Measures 78
4.3 Simulation study 79
4.3.1 Model for the coalescent with fertility correlation 79
4.3.2 Impact of fertility transmission on tree imbalance 80
4.3.3 Robustness of the method 80
4.3.4 The effect of tree reconstruction 82
4.3.5 Experimental data 82
4.4 Results 83
4.4.1 Power of tests based on tree imbalance 83
4.4.2 Robustness of the method 84
4.5 Discussion 90

5 Minimal clade size and external branch length under the neutral coalescent 97

5.1 Introduction 98
5.2 Background and Notations 99
5.3 Main results 100
5.4 The size of the minimal clade 105
5.4.1 Level of coalescence with the rest of the sample 105
5.4.2 Minimal clade size 107
5.5 Some comparisons 110
5.5.1 Two random genes 110
5.5.2 A random clade 112
5.6 External branch lengths 114
5.6.1 Unconditional distribution 114
5.6.2 Conditional distributions 115
5.6.3 Application: multilocus haplotypes 120

6 Brownian Models and Coalescent Structures 123

6.1 Introduction 124

6.2 Models 126
6.2.1 Coalescent trees 126
6.2.2 Random walks 127
6.2.3 Brownian motion as an approximate model of stepwise mutation 128
6.3 Estimation based on pairwise statistics 130
6.3.1 Basic results about X_n 130
6.3.2 Squared distances 130
6.4 Estimation based on likelihood 133
6.4.1 A peeling algorithm 133
6.4.2 Pseudomaximum-likelihood Algorithm 136
6.4.3 Lower bound of the variance of estimators of θ 141
6.4.4 Markov chain Monte Carlo 143
6.5 Spatial dispersal: Application to a biological dataset 146
6.6 Discussion 148

7 Conclusion **151**
7.1 English Conclusion 155

Bibliographie **157**

A Convergence of compound Poisson process toward Brownian motion **175**

Chapitre 1

Introduction

La présente thèse vise à appliquer des concepts de probabilités et de statistique aux domaines connexes que sont la systématique biologique et la génétique des populations. La systématique cherche à classer les espèces, à établir des relations entre elles, en fonction de leur histoire évolutive. La génétique des populations opère à un niveau plus microscopique et s'intéresse à la diversité génétique de populations d'êtres vivants. La structure de généalogie joue un rôle central dans ces deux disciplines, elle sera notre fil directeur.

En systématique, elle est appelée phylogénie (du grec phylon : race et geneia : origine) et représente l'histoire évolutive des espèces. Bien que le fait de relier les espèces en fonction de leurs degrés d'affiliation ait été introduit avant Darwin (voir Panchen, 1992), c'est lui qui donna un sens évolutif à cette représentation (Darwin, 1859, p.117). En génétique des populations, une généalogie représente l'histoire commune d'individus ou de gènes et peut être représentée par un modèle mathématique appelé le coalescent (Kingman, 1982a). C'est une dénomination imagée : en adoptant une vision rétrospective du temps, deux gènes coalescent, autrement dit fusionnent, dès l'instant où ils ont un ancêtre commun. La confusion entre coalescent et phylogénie est entretenue par le fait que les phylogénies sont souvent estimées à partir de données génétiques. Il est bon de rappeler que les gènes ne sont qu'un moyen de retrouver une phylogénie. Pour simplifier, il n'existe qu'une seule phylogénie reliant un ensemble d'espèces et il est souhaitable que différents gènes permettent de retrouver la même phylogénie. Au contraire, au

sein d'une population, chaque gène est susceptible d'avoir une histoire, i.e. un coalescent, différente (voir Figure 1.1).

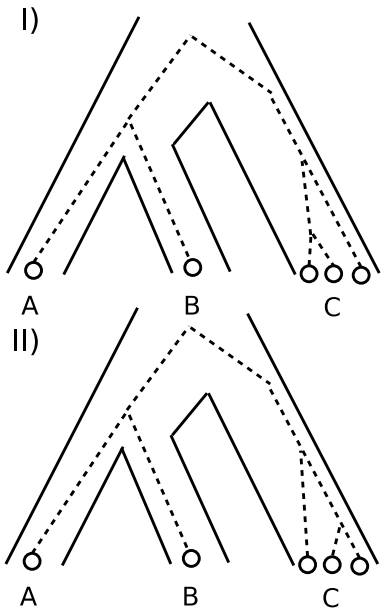

FIG. 1.1 – La phylogénie des trois espèces A, B et C est représentée avec les traits continus. Les généalogies du gène N°1 (I) et N°2 (II) sont représentées en pointillé. Les 3 individus de l'espèce C n'ont pas le même coalescent suivant que l'on considère le gène N°1 ou le gène N°2. Par contre, les deux généalogies de gènes permettent de retrouver la vraie phylogénie.

Dans la suite, nous emploierons le mot généalogie pour désigner indifféremment un coalescent ou une phylogénie. Suivant la discipline, ces généalogies sont introduites à des fins différentes. La cladistique, attribuée à Hennig (1966), classe les espèces en fonction de leurs liens de parenté dans les phylogénies. Elles sont le support pour reconstituer l'histoire évolutive de caractères morphologiques ou physiologiques. Au contraire, en génétique des populations, les généalogies de gènes ne sont que rarement intéressantes en tant que telles. Elles sont considérées comme une donnée manquante qu'il est souhaitable de reconstruire afin de mieux connaître la démographie ou la génétique d'une population. Toutefois cer-

taines généalogies intra-spécifiques ont suscité un grand intérêt, notamment celle reliant l'ADN mitochondrial des populations humaines (Cann, Stoneking et Wilson, 1989). L' interprétation de ces généalogies intra-spécifiques reste difficile et est sujet à controverse (Excoffier et Langaney, 1989). Bien que l'étude de ces généalogies réponde à des motivations différentes, elles ont en commun le fait de contenir de l'information biologique. C'est plus particulièrement l'information contenue dans la forme de ces généalogies qui nous occupera par la suite.

1.1 Modèles stochastiques en phylogenèse

La primeur de l'étude de la forme des phylogénies est traditionnellement attribuée à un groupe de paléontologistes et biologistes des populations qui se réunirent à Woods Hole au début des années 70 (Raup et al., 1973 ; Gould et al., 1977 ; Schopf, 1979). Ils souhaitaient montrer que des processus stochastiques pouvaient expliquer des phénomènes macroévolutifs. Jusqu'à cet instant, les explications étaient de nature déterministe. Par exemple, la surreprésentation d'espèces de lézards iguanidés appartenant aux 4 genres *Anolis, Chamaeleolis, Chamaelinorops* et *Phenacosaurus*, était expliquée par l'apparition d'un « tampon adhésif sous la peau ». Cette innovation est considérée comme une adaptation morphologique qui leur permit de conquérir des niches écologiques (Patterson, 1983). Au contraire, Gould (1977) essaya de donner des explications qui n'étaient ni « taxon-spécifiques »comme l'est l'exemple précédent, ni « temps-spécifiques ». L'exemple d'explication « temps-spécifique »donnée par Gould (1977) est celui des espèces de mammifères, peu nombreuses jusqu'au tertiaire en raison de la concurrence des dinosaures, qui explosèrent après la disparition de ces derniers. Le groupe de Woods Hole mit en avant la part du hasard pour expliquer les variations du nombre d'espèces au cours du temps. Cette volonté est comparable à celle des « neutralistes », en génétique des populations, qui mirent l'accent sur la dérive génétique, phénomène stochastique, comme facteur majeur de l'évolution des fréquences géniques (voir Millstein, 2000). Gould (1977, p.32) est assez explicite à ce sujet : « How different then, is the world from the stochastic sytem ? [...] The answer would seem to be "not very". »

L'introduction de modèles stochastiques en phylogénie remonte en fait au début du XX^{e} siècle. Willis (1922) avait remarqué l'aspect concave de l'histogramme du nombre d'espèces par genre. La majorité des genres compte très peu d'espèces tandis que quelques genres comptent un nombre d'espèces très important. L'exemple le plus classique est celui du nombre très important d'espèces de scarabées parmi les insectes qui avait fait dire au célèbre biologiste J. B. S. Haldane que Dieu avait une affection démesurée pour les scarabées. Yule (1924) introduisit un processus de branchement à taux constant (aussi appelé processus de naissance pure à taux constant) où les événements de branchement correspondent aux spéciations. Dans l'intervalle $[t, t + dt]$, la probabilité qu'une espèce donne naissance à une nouvelle espèce est égale à $\lambda dt + o(dt)$. De manière indépendante, chaque genre donne naissance, avec un taux constant μ à une nouvelle espèce qui sera la première d'un nouveau genre. Sous les hypothèses du modèle de Yule, la loi du nombre d'espèces par genre $P(N = n)$ est une loi puissance qui décroît comme $n^{-(1+\mu/\lambda)}$ (Aldous, 2001). En adaptant les valeurs des paramètres aux données, Yule trouva une très bonne adéquation avec les mesures de Willis. En géneral, les auteurs ne considèrent, dans le modèle de Yule, que la partie correspondante à la genèse des espèces et pas celle ayant trait à la genèse des genres. Dans ce cas précis, le modèle de Yule à n espèces est un processus de naissance pure à temps continu qui commence avec une lignée. Chaque lignée a une durée de vie qui suit une loi exponentielle de paramètre λ, et se sépare à la fin de sa vie en 2 lignées (une lignée correspond à l'espèce ancestrale et l'autre à la nouvelle espèce). Le processus continue jusqu'à ce que n lignées aient été formées. C'est cette définition du modèle de Yule que nous utiliserons par la suite. Le modèle de Yule est aussi celui qui a été étudié par le groupe de Woods Hole. Par tradition, il est considéré comme le modèle nul de la phylogenèse.

Afin d'étudier les différentes structures de la diversité des êtres vivants, les biologistes ne se limitent plus à des données taxonomiques comme le nombre d'espèces par genre mais considèrent désormais les phylogénies. Willis qui se demandait pourquoi la répartition du nombre d'espèces par genre était si hétérogène s'interrogerait aujourd'hui sur l'aspect déséquilibré qu'ont la majorité des phylogénies publiées. Par déséquilibre, nous entendons le fait que les espèces ancestrales (les noeuds internes) ont un nombre de descendants (les feuilles) inégales de part

et d'autre des deux sous-arbres dont ils sont racines (voir Mooers and Heard, 1997 pour une revue très complète sur la forme des phylogénies). Les héritiers du groupe de Woods Hole se sont demandés dans quelle mesure le déséquilibre des phylogénies publiées était compatible avec le modèle de Yule. En général, ils trouvent que le modèle de Yule prédit des arbres plus équilibrés que ceux observés (Guyer et Slowinski 1991, 1993, Heard, 1992, Mooers, 1995, Aldous, 2001) à l'exception de Savage (1983) qui étudia cependant des phylogénies comprenant au plus 7 espèces. Un grand nombre d'indices a été introduit afin de mesurer ce déséquilibre (Kirkpatrick et Slatkin, 1993 ; Agapow and Purvis, 2002). Des modèles alternatifs au modèle de Yule prenant en compte des différentiels de taux de sélection entre lignées (Kirkpatrick et Slatkin, 1993), des hypothèses du type « découverte de niche écologique »(McKenzie et Steel, 2001), des corrélations entre le taux de spéciation et un trait biologique (Heard, 1996) ont été proposés et leurs effets sur les mesures de déséquilibre ont été étudiés à partir de simulations de Monte-Carlo.

D'autres disciplines émanant de la systématique ont considéré un modèle stochastique différent du modèle de Yule. Ils ont considéré que toutes les phylogénies à n espèces étaient équiprobables. En biogéographie, les phylogénies d'espèces appartenant aux mêmes zones géographiques sont utilisées pour retrouver des événements biogéographiques du passé (Hennig, 1966). Afin de quantifier dans quelle mesure les phylogénies sont congruentes, la probabilité que deux phylogénies soient congruentes par hasard a été calculée (Rosen, 1978). Néanmoins, Simberlof et al. (1981) ont critiqué le modèle uniforme qui semble dénué de motivations biologiques. Ces probabilités de congruence ont été calculées, par la suite, dans le cadre du modèle de Yule (Brown, 1994). Une approche similaire a été envisagée dans le cadre de l'interaction hôte-parasite afin de savoir si les deux types d'espèces avaient évolué de concert (Page, 1990).

1.2 Modèles stochastiques en génétique des populations

Tandis que la place du hasard est encore modeste en phylogenèse (mise à part les modèles de mutations génétiques), elle est prépondérante en génétique des populations depuis l'introduction par Wright et Fisher, au milieu du XX^{e} siècle d'un modèle prenant en compte les différences de succès reproducteur chez les individus (Fisher, 1930 ; Wright, 1931). Cette approche à temps discret met en évidence la dérive génétique, i.e. la fluctuation aléatoire des fréquences géniques au sein d'une population. Kimura (voir Watterson, 1996 pour un récapitulatif de ces travaux) étudia nombre de propriétés de ce modèle (comme le temps de fixation d'un allèle avec éventuellement des mutations, de la sélection, de la migration ...) en utilisant une approximation de diffusion (voir Wakeley, 2005 pour une revue orientée mathématique de l'approximation de la diffusion en génétique des populations). Des mathématiciens et des biologistes (Felsenstein, 1971 ; Griffiths, 1980) commencèrent à adopter une approche rétrograde en étudiant les propriétés des généalogies issues du modèle de Wright-Fisher. C'est Kingman (1982a,b) qui formalisa cette approche en étudiant à la fois la loi des temps de la généalogie mais aussi celle de sa topologie. Bien que la découverte du coalescent soit attribuée à Kingman (on parle même de coalescent de Kingman), un de ses contemporains introduisit le même objet (Tajima, 1983) de manière indépendante. D'ailleurs, les biologistes qui ont une approche à temps discret du coalescent et non à temps continu comme Kingman, utilisent plutôt la description de Tajima.

Le coalescent diffère des modèles classiques de la génétique des populations, non seulement parce qu'une vision rétrograde du temps est adoptée mais aussi parce qu'il ne traite que d'un échantillon d'individu et non de toute la population (voir Figure 1.2). L'explosion récente des techniques de biologie moléculaire a ainsi permis de récolter de nombreuses données de polymorphisme génétique (typage d'individus issus d'une ou plusieurs populations pour un ou plusieurs loci polymorphes) dont le traitement statistique est effectué dans le cadre de la coalescence. La manière dont la communauté des généticiens des populations s'est emparée du coalescent peut-être divisée en 3 parties

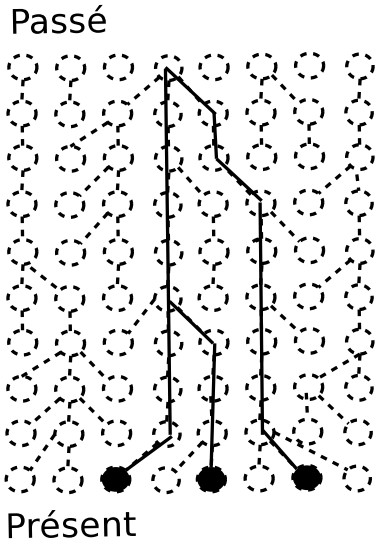

FIG. 1.2 – Une réalisation du modèle de Wright-Fisher haploïde à 8 individus pendant 10 générations. Chaque cercle représente un individu et les traits en pointillé représentent les relations parent-enfant. L'arbre de coalescence de 3 individus est représenté en trait gras.

- Une approche théorique qui a permis de retrouver, de manière plus simple, des résultats connus par l'approximation de la diffusion.

  On peut citer, entre autres, le calcul de l'espérance de l'estimateur du taux de mutation introduit par Watterson (1975). Cet estimateur compte, dans un échantillon de n séquences, le nombre de nucléotides où est survenu une mutation. On peut aussi citer le résultat récent de Griffiths (2003) qui trouva l'âge moyen d'un allèle porté par j individus dans un échantillon de taille n tandis que Kimura trouva l'age moyen d'un allèle porté par une fraction x de la population toute entière.
- Une manière rapide de simuler des marqueurs génétiques pour un échantillon d'individus.

  Les simulations sous le modèle de Wright-Fisher sont coûteuses puisqu'elles nécessitent la prise en compte de la population toute entière et requièrent un tirage aléatoire à chaque génération. Au contraire, le coalescent ne traite que de l'échantillon de n individus et, dans les cas simples, le tirage aléatoire de $n-1$ temps de coalescence. Des logiciels de simulation se sont multipliés, on

peut citer, entre autres, SIMCOAL2 (Excoffier, 2000) qui prend en compte de nombreux aspects de la démographie d'une population, différents modèles de mutation génétique ainsi que les événements de recombinaison.

– Un modèle statistique.

  Une approche reposant sur le calcul de la vraisemblance s'est développée afin d'estimer les paramètres démographiques et génétiques des modèles d'une part, et de faire de la sélection de modèles d'autre part (Nielsen, 2001). La vraisemblance s'exprime comme une intégrale sur l'ensemble des coalescents. Cette intégrale ne peut se calculer de manière exacte et est estimée par une méthode de Monte-Carlo. Les coalescents sont simulés suivant une loi d'importance (Kuhner, 1995) en utilisant une méthode de Monte-Carlo par chaîne de Markov (MCMC).

La sensibilité de la forme des arbres de coalescence à la démographie a été largement étudiée. Par exemple, dans les populations qui ont vécu une forte expansion démographique, les généalogies de gènes auront une forme étoilée (Di Rienzo et Wilson, 1991). Autrement dit les gènes coalesceront à l'époque où la taille de la population était faible. En revanche, le déséquilibre des arbres de coalescence n'a pas suscité beaucoup d'intérêt jusqu'à présent. Un indice qui mesure de manière très incomplète le déséquilibre a été utilisé à plusieurs reprise. C'est la fréquence avec laquelle une branche externe, i.e. une branche directement connectée aux feuilles de l'arbre, se connecte à la racine. Przeworski (1999) montra que la sélection purificatrice (modèle où les mutations diminuent la valeur sélective d'un individu) ne modifiait pas sensiblement cet indice. Récemment, Maia et al. (2004) confirmèrent ce résultat, en utilisant, cette fois-ci, les indices de déséquilibre introduits en phylogénie. Au contraire, Sibert (2002) montra que les branches externes se connectaient plus souvent à la racine lorsque la fertilité se transmettait partiellement d'une génération à une autre. Barton (1998) ansi que Fay et Wu (2000) suggérèrent, sans l'étayer pour autant par des simulations, que le balayage sélectif (des allèles avantagés par la sélection se fixent rapidement) pouvait déséquilibrer les généalogies à des loci proches de celui soumis à sélection. Krings et al. (1997), après avoir reconstruit la généalogie d'ADN mitochondrial appartenant à 986 humains contemporains et un homme de Néanderthal, remarquèrent que la branche du Néanderthal se connectait directement à la racine. Nordborg (1998) précisa

dans quelle mesure cette observation pouvait prouver que les deux espèces étaient bien distinctes.

1.3 Présentation des chapitres

Nous commençons par introduire quelques notations qui sont illustrées par la figure 1.3. Les relations de parenté entre espèces sont représentées par des arbres binaires phylogénétiques enracinés. De tels arbres contiennent des noeuds de degré 1, 2 ou 3 (un unique noeud de degré 2). Par degré d'un noeud, nous entendons le nombre de branches aboutissant à ce noeud. Les noeuds de degré 1 sont étiquetés et appelés feuilles. Les noeuds de degré 2 ou 3 sont appelés noeuds internes et ne sont pas étiquetés. Le noeud de degré 2 a un statut particulier : c'est la racine de l'arbre. Une branche externe connecte une feuille à un noeud interne tandis qu'une branche interne connecte deux noeuds internes entre eux. Les arbres binaires phylogénétiques enracinés qui ne contiennent pas d'échelle temporelle (comme celui de la figure 1.3) sont parfois appelés cladogrammes (voir Aldous, 1995). L'ordre d'étiquetage des feuilles a bien sûr son importance : le cladogramme obtenu à partir de celui de la figure 1.3 en intervertissant l'étiquette A et l'étiquette D est différent du cladogramme initial. En revanche, aucune distinction n'est faite entre les branches gauches et les branches droites de l'arbre : le cladogramme obtenu à partir de celui de la figure 1.3 en intervertissant l'étiquette A et l'étiquette B est identique au cladogramme initial

La suite du manuscrit est structurée de la manière suivante. Le chapitre 2 est consacré à l'étude d'indices sensibles au déséquilibre d'une généalogie et à la puissance de ces indices à rejeter le modèle nul. Dans le chapitre 3, nous étudions précisément l'asymptotique de deux statistiques mesurant le déséquilibre d'une phylogénie. Les propriétés des statistiques sont étudiées dans le modèle de Yule ainsi que dans le modèle uniforme qui suppose que toutes les phylogénies à n feuilles sont équiprobables. Le chapitre 4 montre comment le déséquilibre des généalogies peut être interprété en génétique des populations. Le chapitre 5 est consacré à l'étude de variables sensibles à la forme d'un arbre de coalescence. Ces variables sont liées à l'événement de coalescence entre un individu et ses

plus proches parents parmi un échantillon d'individus. Le chapitre 6 introduit différentes méthodes d'inférence d'un paramètre d'intérêt en génétique spatiale.

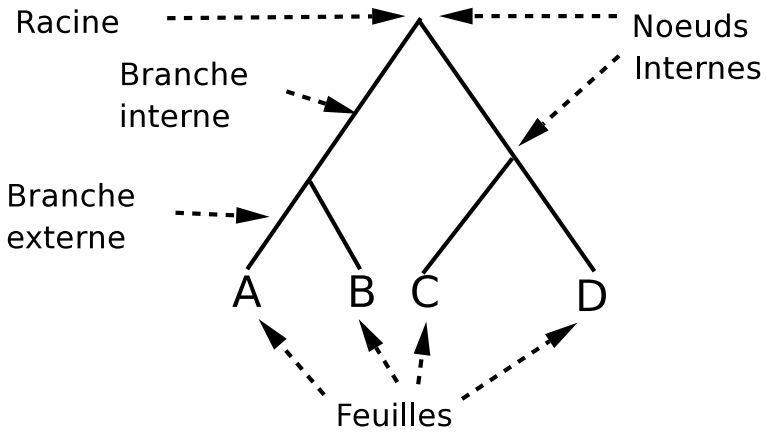

FIG. 1.3 – Notations associées aux arbres binaires phylogénétiques

L'indice de Sackin et consorts revisités

Nous avons déjà mentionné que différentes statistiques avaient été introduites pour mesurer le déséquilibre d'une phylogénie. Les distributions asymptotiques de ces statistiques n'étaient jamais connues, exceptés pour la statistique comptant le nombre de sous-arbres comprenant 2 feuilles (McKenzie et Steel, 2000), appelés de manière imagée les cerises (Figure 1.4). Plutôt que de se limiter au nombre de cerises, nous avons étudié la famille de statistiques comptant le nombre de sous-arbres à x feuilles dans un arbre à n feuilles ($2 \leq x \leq n-1$). Nous rappelons que la notion de sous-arbre peut être définie à partir de la relation d'ordre « descendre de »qui est naturellement induite par un arbre $\mathcal{A}$: le noeud x descend du noeud y si et seulement si y appartient au chemin allant de x à la racine (en particulier x descend de x). Le sous-arbre de l'arbre $\mathcal{A}$ issu d'un noeud x contient tous les descendants de x et l'ensemble des sous-arbres de $\mathcal{A}$ contient la réunion des sous-arbres issus de x où x parcourt l'ensemble des noeuds. En adaptant des résultats de Devroye (1991), nous avons étendu le théorème central-limite obtenu par Mc Kenzie et Steel (2000) : nous avons démontré que les statistiques dénombrant les sous-arbres à x feuilles étaient asymptotiquement gaussiennes dans le

modèle de Yule. Au lieu de s'intéresser à toutes ces statistiques séparément, un indice introduit par Sackin (1972) prend en compte toutes les informations contenues dans celles-ci. Nous avons montré que la loi limite de cet indice est la même que celle du nombre de comparaisons effectuées par le célèbre algorithme de tri rapide : le Quicksort (Hoare, 1962). Cette observation, apparemment incongrue, est justifiée par le fait que le modèle de Yule est intimement lié au modèle des permutations uniformes (dont la définition est donnée dans le paragraphe suivant) introduit pour l'analyse probabiliste du tri rapide. La puissance de ces différentes statistiques a été estimée par des simulations de Monte-Carlo. Le modèle alternatif proposé par Kirkpatrick et Slatkin (1993) suppose qu'une espèce de taux de spéciation r, donne naissance, après un événement de spéciation à une espèce de taux de spéciation pr et à une autre de taux $(1-p)r$ $(0 < p < 1)$. De manière assez prévisible, c'est l'indice de Sackin qui est de puissance maximale tandis que le nombre de cerises est tout de même le meilleur indice parmi toutes les statistiques comptant le nombre de sous-arbres de taille x.

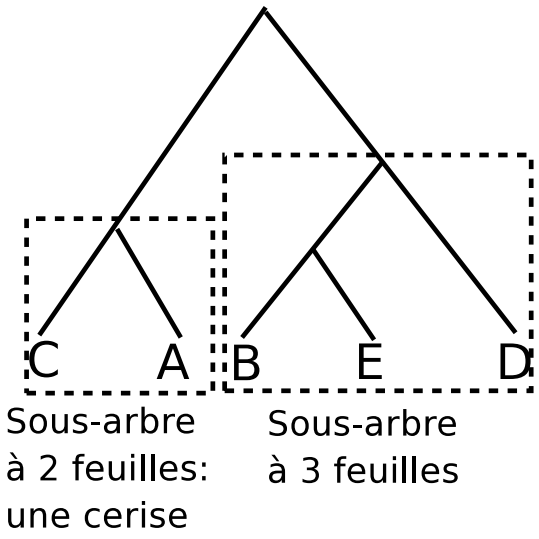

FIG. 1.4 – Cerises et sous-arbres

Lois limites de l'indice de Colless et de Sackin

Le lien entre les modèles probabilistes d'arbres aléatoires en phylogénie et en analyse des algorithmes n'avait pas été assez souligné, mis à part par Aldous (1996, 2001). Nous rappelons dans ce paragraphe quelques définitions et modèles propres à l'analyse des algorithmes. Une définition récursive des arbres binaires tels qu'ils sont utilisés dans cette discipline est donnée par Sedgewick et Flajolet (1996) : « Un arbre binaire est soit un noeud externe, soit un noeud interne auquel on attache une paire ordonnée d'arbres binaires appelés sous-arbre gauche et sous-arbre droit ». Notons que les feuilles sont appelées, dans cette définition, noeuds externes. Il y a deux différences entre les arbres binaires tels qu'ils sont définis précédemment et ceux étudiés en phylogénies (les cladogrammes) : les feuilles des arbres binaires ne sont pas étiquetées tandis que celles des cladogrammes le sont et la paire constituée du sous-arbre droit et du sous-arbre gauche est ordonnée tandis que celle d'un cladogramme ne l'est pas. Le nombre d'arbres binaires à n noeuds internes et $n+1$ feuilles est donné par le nombre de Catalan $\mathcal{C}_n$

$$\mathcal{C}_n = \frac{\begin{pmatrix} 2n \\ n \end{pmatrix}}{n+1}.$$

Une utilisation classique des arbres binaires en informatique repose sur l'algorithme de l'arbre binaire de recherche. Un arbre binaire de recherche est un arbre binaire dans lequel des clés sont associées aux noeuds internes. Ces clés vérifient la contrainte suivante : la clé de tout noeud est supérieur ou égale à toutes les clés de son sous-arbre gauche, et inférieure ou égale à toutes les clés de son sous-arbre droit. Le modèle probabiliste d'arbre binaire le plus classique en analyse des algorithmes est d'ailleurs lié à cette structure d'arbre binaire de recherche. C'est le modèle des permutations uniformes : à partir d'une permutation tirée de manière uniforme parmi l'ensemble des permutations des entiers 1 à n, un arbre binaire de recherche à n noeuds internes est construit. Le premier élément de la permutation est inséré à la racine, le second à gauche de la racine s'il est inférieur à la racine ou à droite dans le cas contraire. La construction de l'arbre se poursuit de la sorte jusqu'à ce que tous les éléments de la permutation aient été insérés

dans l'arbre (voir Figure 1.5). L'autre modèle d'arbre binaire qui suscite l'intérêt des algorithmiciens est le modèle de Catalan, aussi appelé modèle uniforme, qui suppose que tous les arbres binaires à n noeuds internes sont équiprobables.

L'intérêt de ce détour par l'analyse des algorithmes provient de la correspondance quasi-bijective entre le modèle de Yule et le modèle des permutations uniformes d'une part et les deux modèles uniformes d'autre part. Cette correspondance sera explicitée dans le chapitre 3. Elle sera la clé pour étudier les lois asymptotiques des indice sensibles au déséquilibre d'une phylogénie puisque nous utiliserons des méthodes apparues récemment dans l'analyse probabiliste des algorithmes.

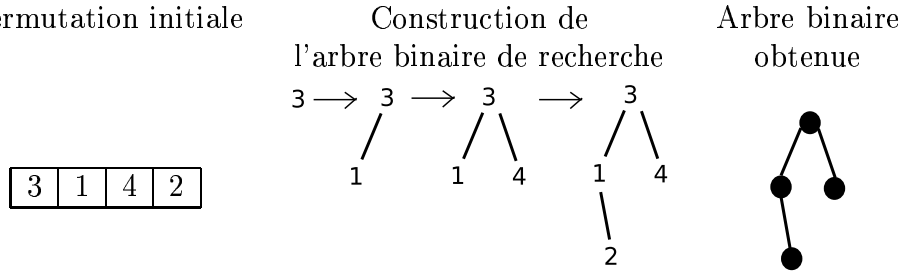

FIG. 1.5 – Construction d'un arbre binaire de recherche dans le modèle des permutations uniformes. Un arbre binaire est construit à partir de la permutation (3,1,4,2).

Les variables aléatoires que nous étudions sont définies par la relation de récurrence

$$S_n \stackrel{d}{=} S_{I_n} + S_{J_n} + t(I_n, J_n)$$

où $\stackrel{d}{=}$ représente l'égalité en loi et I_n (resp. J_n) le nombre de feuilles du sous-arbre gauche (resp. droit) avec $I_n + J_n = n$. Dans le cas du modèle de Yule ou du coalescent de Kingman, I_n suit une loi uniforme sur $\{1 \dots n-1\}$. Par analogie avec l'analyse des algorithmes, nous appellerons $t(I_n, J_n)$ la fonction coût. Les statistiques de Sackin et de Colless vérifient l'équation de récurrence précédente et ont respectivement pour fonction coût $t(I_n, J_n) = I_n + J_n = n$ et $t(I_n, J_n) = |I_n - J_n|$. Dans le modèle de Yule et le modèle uniforme, nous avons établi la convergence de la loi jointe des statistiques ainsi que la convergence des moments (seulement ceux d'ordre 2 dans le modèle de Yule). Dans le modèle de Yule, la loi

limite est caractérisée par une équation de point fixe. La démonstration repose sur la méthode de contraction introduite par Rösler (1991) afin d'étudier le nombre de comparaisons effectuées par le Quicksort. Dans le modèle uniforme, l'étude des propriétés de l'arbre est réduite à celle d'une excursion discrète qui lui est associée (Takacs, 1991). Les lois limites peuvent alors s'exprimer en fonction de l'excursion brownienne.

Les propriétés asymptotiques de ces statistiques avaient déjà été étudiées par Rogers (1994, 1996) à partir de calculs numériques. Il avait remarqué qu'elles étaient très corrélées dans les deux modèles. La corrélation semblait converger vers une valeur proche de 0.98 dans le modèle de Yule et vers 1 dans le modèle uniforme (Rogers, 1996, Fig. 5b). En calculant quelques valeurs du moment d'ordre 3 de la statistique de Colless (Rogers, 1994, Fig. 4), il avait suggéré que cette statistique ne convergeait pas vers une gaussienne. Tous ces résultats s'avèrent être exacts et nous avons pu les établir rigoureusement.

Déséquilibre des généalogies et transmission de la fertilité

Le déséquilibre des arbres binaires est le plus souvent étudié dans un cadre inter-spécifique. Il n'avait attiré que peu d'attention dans le cadre intra-spécifique propre à la génétique des populations. Nous nous intéressons plus particulièrement à l'effet d'un phénomène démographique, la transmission de la fertilité (Heyer et al., 2005), sur ce déséquilibre. Dans les populations où la fertilité se transmet, les individus dont les parents ont eu beaucoup d'enfants vont avoir tendance à plus se reproduire à leur tour. Le modèle que nous étudions à été introduit par Sibert et al. (2002) afin d'étendre le modèle de Wright-Fisher. Dans une population haploïde de taille N, le modèle de Wright-Fisher peut être vu comme un modèle d'urne avec remise. A une génération donnée, l'urne contient les N parents, ils tous la même probabilité $1/N$ d'être choisi par un enfant. Dans le modèle de Sibert et al. (2002), cette probabilité n'est plus la même pour tous. Le parent potentiel i sera choisi avec une probabilité proportionnelle à s_i^α où s_i correspond à la taille de la fratrie du parent potentiel i et α représente l'intensité de la transmission de la fertilité. Plus cette intensité est forte, plus les généalogies d'individus seront déséquilibrées. Nous avons montré que cette observation était robuste en considérant des populations

à taille variable où la fertilité ne se transmettait que pendant un nombre restreint de générations.

D'autres phénomènes démographiques sont aussi susceptibles d'influencer la forme des généalogies. C'est pourquoi nous avons étudié l'effet de la structuration géographique sur le déséquilibre. Suivant les scénarios de structuration envisagés, les généalogies de gènes peuvent être déséquilibrées. Toutefois, la proportion de généalogies déséquilibrées reste toujours beaucoup plus faible que dans le cas où la fertilité se transmet. Les méthodes de reconstruction de généalogies peuvent aussi influencer ce déséquilibre. Cet aspect est important puisque le déséquilibre ne sera jamais calculé à partir de la vraie généalogie mais simplement à partir d'une reconstruction. Nous avons montré que ces méthodes avaient une tendance faiblement prononcée à rendre les généalogies plus asymétriques.

Afin de tester si la fertilité se transmettait effectivement dans certaines populations humaines, nous avons analysé la base de données MOUSE (Mitochondrial and Other Useful SEquences). Elle contient des séquences du gène HV1 (Hyper-Variable N°1), présent dans une partie de l'ADN mitochondrial (la Dloop) réputée pour avoir un fort taux de mutation. Nous nous sommes restreints à l'étude des populations qui comptaient plus de 50 individus échantillonnés. L'ADN mitochondrial se transmettant uniquement par la mère, c'est la composante maternelle de la transmission de la fertilité que nous avons ainsi pu détecter. Les généalogies des 41 populations étudiées ont ainsi été reconstruites et leur déséquilibre calculés. Ce sont les généalogies issues des populations traditionnelles, comme les chasseurs-cueilleurs, qui étaient les plus déséquilibrées. Cette observation suggère une plus forte transmission de la fertilité dans ces populations.

Taille du clade minimal et longueurs des branches externes du coalescent

Dans un coalescent comprenant n individus, nous étudions les lois des variables aléatoires liées à l'événement de coalescence entre un individu et les $n-1$ individus restants (voir Figure 1.6). La première variable aléatoire étudiée est la taille du clade minimal. Un clade minimal contient l'individu de référence ainsi que tous ses plus proches parents. Cette variable aléatoire est simplement liée à la loi de

la topologie du coalescent. Comme la loi de la topologie du coalescent est la même que celle du modèle de Yule, sa distribution reste donc la même dans le cadre du modèle de Yule. La seconde variable aléatoire, en revanche, prend en compte l'aspect temporel. Elle mesure le temps qui sépare cet individu de référence de l'ancêtre qu'il partage avec ses plus proches parents. Cette variable aléatoire correspond aux longueurs des branches externes de l'arbre de coalescence.

La taille du clade minimal suit une loi puissance de paramètre 3. Cette distribution peut-être comparée à celle d'un clade pris au hasard dans la population. Par clade tiré au hasard, nous entendons l'ensemble des individus descendants d'un noeud interne tiré de manière uniforme parmi les $n-1$ noeuds internes. Cette dernière loi suit une loi puissance de paramètre 2. Un clade minimal aura donc donc tendance à compter moins d'individus qu'un clade tiré au hasard. L'espérance du clade minimal reste bornée tandis que celle du clade aléatoire diverge. Au chapitre 1, nous avons établi que le nombre de clades (alors appelés sous-arbres) de taille x était asymptotiquement gaussien. Dans ce chapitre, nous montrons que ce résultat reste vrai pour le nombre de clades minimaux de taille x.

Quant à la loi d'une branche externe de l'arbre, nous montrons que c'est un mélange de loi exponentielle. Ce résultat n'est pas surprenant puisque les temps inter-coalescence suivent des lois exponentielles. Connaissant la loi des branches externes, nous pouvons calculer celle du nombre de substitutions Δ entre un gène de l'individu de référence et celui d'un de ses plus proches parents. Le modèle de mutation envisagé est le modèle des sites infinis (Waterson, 1975) où chaque mutation affecte un nucléotide différent. La loi de Δ est un mélange de lois géométriques décalées (loi de Pascal). Une nouvelle fois, ce résultat n'est pas surprenant, puisque la loi du nombre de substitutions entre deux gènes pris aux hasard dans la population est une loi géométrique décalée.

Ces résultats permettent de quantifier le nombre de loci à utiliser afin d'identifier un coupable parmi une liste de suspects. Si l'ADN du coupable est connu, après avoir relevé par exemple ces empreintes sur les lieux du crime, ainsi que celui d'une liste de suspect, les soupçons se tourneront bien évidemment vers l'individu (le suspect N°1) qui partagent les allèles du coupable. Si le nombre de loci utilisé est insuffisant, il n'est pas garanti que le coupable et le suspect N°1 soient une seule

et même personne. Nous donnons la loi du nombre de générations qui séparent le coupable et le suspect N°1 en fonction du nombre de loci utilisés. Nos résultats étendent ceux précédemment obtenus par Walsh (1992) qui ne considérait pas une liste de suspects potentiels mais un unique suspect.

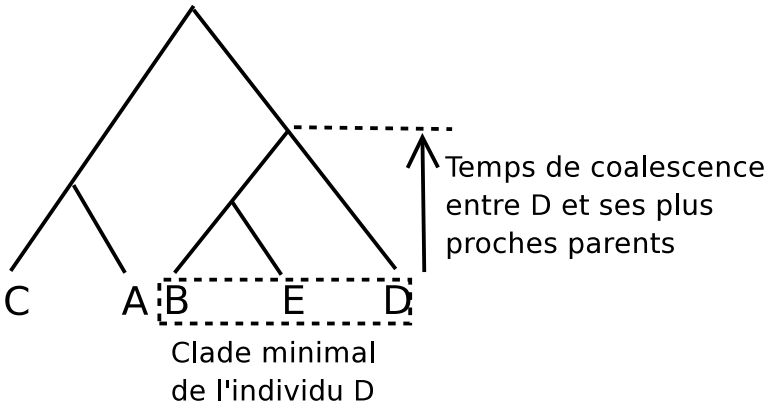

FIG. 1.6 – Clade minimal de l'individu D et temps de coalescence entre D et ses plus proches parents

Coalescent et mouvement brownien

Dans ce chapitre, nous introduisons un modèle qui essaye de rendre compte de la diversité spatiale d'une population. La démarche que nous suivons est inspirée des modèles classiques de mutation en génétique des populations. Les données de gènes au sein d'une population ne peuvent être considérées comme indépendantes puisqu'elles partagent une généalogie commune appelée le coalescent. Suivant le type des données observées, le processus de mutation considéré sera différent. Mais dans tous ces modèles, les événements de mutations surviennent selon un processus de Poisson le long des branches du coalescent. Les données considérées ici ne sont pas génétiques mais géographiques. Nous modélisons l'évolution des positions spatiales au cours du temps par des mouvements browniens qui se propagent le long des branches du coalescent. La structure au second ordre du mouvement brownien considéré vérifie

$$E[B_t^2] = \theta t, \quad t > 0$$

où θ est le paramètre que nous cherchons à inférer. Ce modèle peut aussi être envisagé comme une approximation du modèle de mutation des microsatellites. Ce sont des marqueurs génétiques où une séquence de nucléotides se répète un certain nombre de fois. Le plus simple des modèles de mutation pour les microsatellites est une marche aléatoire sur le nombre de répétitions de la séquence. Une mutation correspond soit à l'insertion d'une séquence soit à une délétion (Kimura et Ohta, 1972). Au cours du temps, l'évolution du nombre de répétitions suit donc un processus de Poisson composé. En le normalisant de manière adéquate, ce processus converge vers le mouvement brownien.

Les propriétés d'un simple estimateur de moment sont étudiées. Cet estimateur calcule la moyenne empirique des carrés des distances entre les individus. Bien que non biaisé, cet estimateur se révèle insuffisant puisqu'il n'est pas consistant. Afin de remédier à ce problème, nous construisons deux estimateurs basés sur le calcul de la vraisemblance. Si la généalogie des individus est connue, la vraisemblance est aisée à calculer à partir d'un algorithme dynamique. Comme cette généalogie est une donnée cachée du modèle, la vraisemblance est obtenue à partir d'une intégrale sur l'ensemble des généalogies possibles. Cette intégrale ne pouvant être calculée de manière exacte, nous l'approchons par une méthode de Monte-Carlo. Afin de ne tenir compte que des généalogies pesant de façon significative sur la vraisemblance, nous avons adopté un échantillonnage d'importance où les généalogies sont simulées suivant la loi conditionnelle $P(G|D)$ où D représente les données et G la généalogie. Les simulations des généalogies suivant cette loi d'importance ont été effectuées avec une méthode de Monte-Carlo par chaîne de Markov.

Nous avons aussi construit un autre estimateur moins coûteux en temps de calcul. La vraisemblance conditionnelle à la généalogie UPGMA (Unweighted Pair Group Method with Arithmetic mean) est calculée. L'estimateur qui maximise cette vraisemblance conditionnelle est biaisé. Des simulations de Monte-Carlo nous ont permis d'exprimer ce biais en fonction de la taille de l'échantillon et l'estimateur a ainsi pu être corrigé. Il est légitime de se demander si ces estimateurs sont efficaces, i.e. sont-ils de variance minimale parmi tous les estimateurs fonctions des données ? Une borne de Cramer-Rao est établie dans le cas où les données ainsi que la généalogie sont connues. A la différence des résultats obtenus

pour le modèle des sites infinis (Tavaré, 2003), la variance des estimateurs obtenus n'approchent jamais cette borne théorique. L' estimateur calculé à partir de la généalogie UPGMA a été appliqué sur un jeu de données réelles. Connaissant les positions géographiques d'ours bruns de Scandinavie, nous avons pu estimer un paramètre de dispersion spatiale. Ce paramètre mesure la distance quadratique moyenne entre une femelle ours et sa mère. La valeur obtenue de 10 km est raisonnable au vu des connaissance biologiques de l'espèce.

1.4 English introduction

In the present work, we apply some concepts of probability and statistics to systematic biology and population genetics. Systematic biology studies the diversity of species and relations between them. Population genetics deals with a more microscopic scale by studying the genetic diversity within a population of organisms. The genealogical structure is of primary importance in both fields and we propose to study some of its properties in the following.

In systematic biology, a genealogy is called a phylogeny. It gives a representation of the links between species. Although it has been studied long before Darwin (see Panchen, 1992), he is the first one to give an evolutionary meaning to it. In population genetics, it represents the common ancestry of a sample of genes or individuals and it is called the coalescent (Kingman, 1982a,b). It is a metaphoric denomination, since genes, looking backwards in time, are said to coalesce when they merge into their common ancestor. Systematicians are interested by the phylogenetic tree in itself. The classification of species is based on phylogenies according to principles of cladistics (Hennig, 1966). On the contrary, the coalescent is not reconstructed in general for its own. It should be seen as a missing data that is useful for inferring the past demography of a population as well as selective processes. Only a few coalescent trees are interesting for their own, such as the one built from mitochondrial DNA of human populations (Cann, Stoneking and Wilson, 1989). Interpreting such intra-specific genealogies should be led with precaution and could be misleading (Excoffier et Langaney, 1989). In any case, although phylogenetic and coalescent trees are introduced for dif-

ferent purposes, they both contain biological information. More specifically, the information contained in the shape of these phylogenies will retain our attention.

Stochastic model in phylogeny

In the beginning of the seventies, a group of paleontologists held meetings in Woods Hole to stress the role of chance in macroevolution (Raup et al., 1973; Gould et al., 1977; Schopf, 1979). They were wondering to what extent stochastic processes could account for the same macroevolution observations than "deterministic" explanations. They raised a debate between proponents of deterministic explanations and proponents of the role of chance (see Millstein, 2000 for a comparison between this debate and the one opposing selectionists and neutralists in population genetics). The model they studied was a continuous-time pure birth process started with one lineage. Each lineage persists for an amount of time that is exponentially distributed (with parameter λ) and then splits into two lineages. The process continues until there are n lineages. According to this model, each species has the same probability $\lambda dt + o(dt)$ to produce a new species during the time interval $[t, t+dt]$. The Woods Hole group tried to assess whether this model can explain the observed variation of the number of species along times. Gould (1977, p.32) was really optimistic about the answer: "How different then, is the world from the stochastic sytem? [...] The answer would seem to be not very."

The introduction of stochastic models in phylogeny dates in fact to the beginning of the twentieth century. Willis (1922) noticed that the distribution of the number of species by genus has a concave shape. Most genus contain just one species whereas a few of them contain many different species. Yule (1924) introduced a branching process with a constant speciation rate. In this model, he found that the number of species per genus follows a power law distribution that provides a very good fit to the data (see Aldous, 2001 for a review of stochastic models for phylogenetic trees). This is the same model that has been studied by the Woods Hole group. It is called the Yule model and it is usually assumed to be the null model in phylogeny.

The counterpart of Willis' observation is that the published phylogenies are unbalanced. An unbalanced tree is a tree where most of the nodes separate the tree

in two subtrees of significatively different sizes (see Mooers and Hears, 1997 for an exhaustive review about tree shape). Followers of the Woods-Hole Group asked whether the shape of published phylogenies is compatible with the Yule model. Generally they found that published phylogenies are by far more unbalanced than expected under the Yule model (Guyer and Slowinski 1991, 1993; Heard, 1992; Mooers, 1995; Aldous, 2001), except Savage (1983) who studied phylogenies with at most 7 taxa. A large number of index sensitive to tree balance has been introduced (Kirkpatrick and Slatkin, 1993; Agapow and Purvis, 2002). The effect of alternative models of speciation on tree shape has been investigated with Monte Carlo simulations (Kirkpatrick and Slatkin, 1993, Heard, 1996, McKenzie and Steel, 2001).

The Yule model is not the only one which has been considered by biologists. The uniform model assuming that all phylogenies with n leaves are equally likely has also been studied in systematic biology. It has been introduced for computing the probability that two phylogenies share common clades by chance (Rosen, 1978). These probabilities have also been computed in the Yule model (Brown, 1984). These probabilities are useful in biogeography. In this field, it is argued that past geographic events would lead similar footprints on different phylogenies of species that live in the same geographic area. The same approach is considered in host-parasite evolution (Page, 1990) in order to test whether the two types of species have coevolved.

Stochastic models in population genetics

While the use of stochastic models for phylogenetic trees is still limited (except for the mutation models), their role has been prevailing in population genetics. It dates to Wright and Fisher (Fisher, 1930; Wright, 1931), who introduced a model that take into account random differences of reproductive success among individuals. Their approach highlights the importance of genetic drift, i.e. the random fluctuation of gene frequencies in a population. Kimura (see Watterson, 1996 for a review of his work) studied many properties of this model (such that the fixation time of an allele) using the approximation of diffusion. Meanwhile, some authors started studying the genealogy of individuals in the Wright-Fisher

model (Felsenstein, 1971; Griffiths, 1980). Kingman (1982 a,b) encompassed these works by studying the time of the genealogy of n sampled individuals as well as its topology. The stochastic process underlying the genealogy was called the coalescent. The same object was studied independently by Tajima (1983).

Two differences between classical models of population genetics and the coalescent shall be highlighted. First, time is counted backwards rather than forward. Secondly, it deals with a sample of n individuals rather than with the whole population containing N individuals. Since the number of molecular markers available has increased dramatically during the past few years, the coalescent provides a formal statistical setting to handle these data. The way population geneticists use the coalescent theory can be divided into three parts.

- A theoretical approach where old results obtained with the approximation of diffusion can be derived in a simpler way.

  We can cite for instance a recent result obtained by Griffiths (2003). On one hand, Griffiths gave the expected age of an allele carried by i individuals in a sample of n individuals. On the other hand, Kimura gave the excepted age of an allele carried by a fraction x of the whole population.

- An efficient way to simulate polymorphism.

  The classical approach to simulate polymorphism is based on the forward-in-time Wright-Fisher model, in which all N individuals have to be simulated. At each generation a random sampling of N individuals has to be performed. Using the coalescent, the genealogy of the n individuals only has to be drawn. It requires $n-1$ random sampling in the simpler cases and the mutations can be superimposed independently on the coalescent tree, as long as neutrality is assumed. For instance, the software SIMCOAL (Excoffier, 2000) simulate genetic markers in a population evolving under various demographic scenarios.

- A statistical framework where the inference of various demographic and genetic parameters can be lead as well as model selection (Nielsen, 2001).

  A wide number of likelihood based methods are driven under the coalescent framework. The likelihood can be expressed as an integral over all the

possible coalescent trees. Since the computations can not be done exactly, Markov chain Monte Carlo methods are generally used in order to sample the most likely genealogies (Kuhner et al., 1995). Importance sampling methods have also been developped for computing the likelihood (Griffiths and Tavaré, 1994b).

However, the imbalance of coalescent trees has not raised as much attention as the imbalance of phylogenies. An index that poorly captures imbalance is sometimes considered in the litterature. It is the number of times an external branch - a branch connected to a leaf - is directly connected to the root. Prezworski et al. (1999) showed that purifying selection does not modify significantly the expected value of this index. Recently Maia et al. (2004) confirmed that purifying selection does not modify tree balance using tree balance indices dedicated to phylogenetic trees. Krings et al. (1997) have reconstructed the genealogy of mtDNA sample from 986 present-days humans and one Neanderthal individual. They noticed that the Neanderthal branch was connected directly to the root suggesting that the two species never interbred. Nordborg (1998) showed nevertheless that this information was not sufficient to prove the existence of two different species.

Thesis contents

We start by introducing some notations which are illustrated in Figure 1.7. Evolutionary relationships are generally represented by rooted binary trees. Such trees contain labelled nodes of degree 1 that are called leaves and unlabelled nodes of degree 2 or 3 that are called internal nodes. The unique node with degree 2 has a special status and it is called the root of the tree. An internal branch connects two internal nodes whereas an external branch connects an internal node with a leaf. Phylogenetic trees without any time scale are called cladograms (see Aldous, 1995). The ordering of the labellings is important, i.e. switching label A and D in Figure 1.7 would lead to a different cladogram. However, there is no distinction between a left branch and a right branch, i.e. switching label A and B would lead to the same cladogram.

The chapter 2 is dedicated to the study of indices capturing tree balance. We establish limiting distributions of these indices and we investigate their power at

detecting a departure from the Yule model. In the chapter 3, we prove rigorously the moment convergence as well as the weak convergence of two widely used statistics that capture tree balance. In the chapter 4 we show that tree balance can be used in population genetics for detecting fertility inheritance. Chapter 5 is devoted to the study of random variables that describe the relatedness of an individual to the rest of the sample in the coalescent. Chapter 6 is dedicated to different inference procedures in spatial genetics.

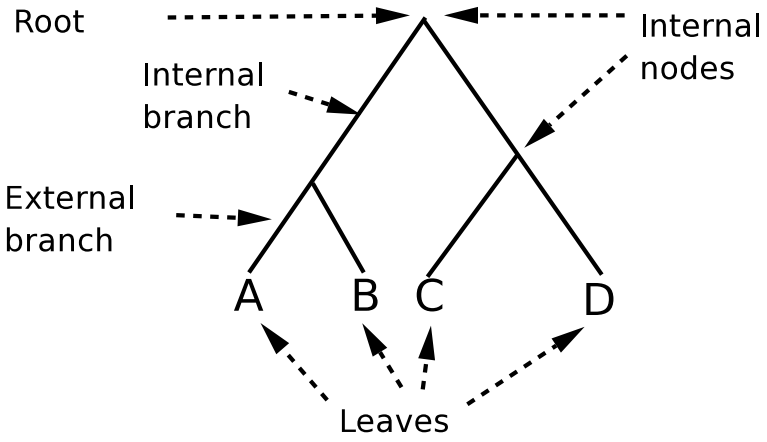

Figure 1.7: Notations associated to binary phylogenetic trees

Chapter 2

On statistical tests of phylogenetic tree imbalance: The Sackin and other indices revisited

Abstract

We investigate the distribution of statistical measures of tree imbalance in large phylogenies. More specifically, we study normalized versions of the Sackin's index and the number of subtrees of given sizes. Using the connection with structures from theoretical computer science, we provide precise description for the limiting distribution under the null hypothesis of Yule trees. Corrected p-values are then computed, and the statistical power of these statistics for testing the Yule model against a model of biased speciation is evaluated from simulations. As an illustration, the tests are applied to the HIV1 reconstructed phylogeny.[1]

[1]Article published: Blum M.G.B. and O. François 2005. On statistical tests of phylogenetic imbalance: the Sackin and other indices revisited. *Math. Biosci.* **195** 141-153.

2.1 Introduction

Phylogenetic trees are widely used in biology to represent evolutionary relationships between species (Nei and Kumar, 2000). A second kind of application is to the study of cladogenesis. In this case, the shape of a phylogenetic tree conveys useful information about the process by which it has grown (Harvey et al, 1996). It may reflect for example the fingerprint of the rates of species formation and extinction. Measuring the degree of imbalance or asymmetry of a tree topology may therefore provide support for the hypothesis that species have different potential for speciation.

Several statistics have been introduced for assessing the level of asymmetry of a tree. These statistics are often used to test whether the tree topology differs significantly from a null model in which the rates of speciation are constant among species (Kirkpatrick and Slatkin, 1993; Mooers and Heard, 1997). The null model is commonly known as the equal-rates Markov model or *Yule model* (Yule, 1924). In the Yule model on rooted trees, each external branch has an equal probability of splitting (Athreya and Ney, 1972).

Among imbalance statistics, the most classical are Sackin's index (Sackin, 1972; Shao and Sokal, 1990) and Colless' index (Colless, 1982). Sackin's index is the average path length from a tip to the root of the tree. Colless' index inspects the internal nodes, partitioning the tips that descend from them into groups of sizes r and s, and computes the sum of absolute values $|r - s|$ for all nodes. Colless's index is often renormalized for giving the value one to the totally pectinate tree. More recently, McKenzie and Steel (2000) proposed to count the number of *cherries*, ie the number of pairs of leaves that are adjacent to a common ancestor node.

The power of five imbalance statistics have been evaluated by Kirkpatrick and Slatkin (1993) who concluded that Sackin and Colless statistics were among the most powerful with respect to an alternative model of biased speciation. These works were extended by Agapow and Purvis (2002) regarding more biologically motivated alternative models. These authors reached similar conclusions.

In this article, we consider fully dichotomous trees (binary trees) with n leaves

(and $(n-1)$ internal nodes). We focus on Sackin's index and its connection to a series of statistics that are similar to the number of cherries. Our emphasis is on the fact that all these measures are relevant to the same empirical distribution, namely the distribution of the size of subtrees. Sackin's index is connected to the expectation of the empirical distribution while cherries merely correspond to subtrees of size two.

The main results presented in this article can be summarized as follows. First, we give a description of the limiting distribution of the Sackin's index for large n. The limiting distribution is non Gaussian, and can be defined as the solution of a functional fixed-point equation. Second, we extend McKenzie and Steel's result for the number of cherries in a Yule tree to the number (or frequencies) of subtrees of size larger than two.

These extensions are based on a link with existing results in theoretical computer science regarding binary search trees. These trees appear as formal representation for *divide-and-conquer* algorithms (Rösler, 1991; Hwang and Neininger, 2002). We exploit the one-to-one correspondence between binary search trees and Yule trees in order to describe the asymptotic distributions of the Sackin's Index and the size of subtrees.

In addition, we propose a new statistic based of the computation of a ℓ_1 distance between the empirical distribution of the number of subtrees and the theoretical distribution. The power of all statistics to reject the null hypothesis of a Yule tree against a model of biased speciation are then evaluated.

The article is organized as follows. Section 2 presents the asymptotic theory for the Sackin's index and the number of subtrees of a given size. Section 3 describes statistical tests with proper corrections based on the theory and their five percent confidence intervals. Section 4 evaluates the statistical power of these tests based on simulation studies. An example of application to the HIV tree is discussed in Section 5.

2.2 Theory

2.2.1 Sackin's statistic

Sackin's statistic is one of the oldest measure that summarizes the shape of a tree (Sackin, 1972; Shao and Sokal, 1990). It adds the number of internal nodes between each leaf of the tree and the root to form the following index

$$S_n = \sum_{i=1}^{n} N_i$$

where the sum runs over the n leaves of the tree and N_i is the number of internal nodes crossed in the path from i to the root (including the root). An equivalent formulation of S_n is by counting the number of leaves under each internal nodes

$$S_n = \sum_{j=1}^{n-1} \tilde{N}_j$$

where $\tilde{N}_j$ is the number of leaves that descend from the ancestor j. This is a well-known result in systematic biology that the expectation of S_n under the Yule model is of order $2n \log n$ (eg, Kirkpatrick and Slatkin, 1993)

$$E[S_n] = 2n \sum_{j=2}^{n} 1/j \, .$$

The variance is more complex, but it can be estimated by noticing the analogy with a classical problem in theoretical computer science. This analogy is a crucial step in defining the proper correction for indices of large phylogenies. Let us explain this briefly. Binary trees are data structures often encountered in computer science, more specifically in connection with *divide and conquer* algorithms. To each Yule tree corresponds a binary search tree in a unique manner (see Aldous, 2001). Using this one-to-one correspondence, Sackin's statistic is equal to the number of comparisons used by the quicksort algorithm to sort a random input (eg, Rösler, 1991). This can also be seen directly because the Sackin's index is

involved in a stochastic recurrence equation

$$S_n = S_J + S_{n-J} + n \tag{2.1}$$

where J is a uniform random variable over the subset $\{1, \ldots, n-1\}$. The recurrence equation is obtained by splitting the tree at the root into two sister clades (the left and right subtrees). Conditional on J, the values S_J and S_{n-J} are independent random variables which correspond to the indices of the left and right subtrees.

Standard computations lead to the result that the variance of S_n is of order $\sigma^2 n^2$ where σ^2 is independent of n (eg, Hwang and Neininger, 2002). In addition, the normalized Sackin's index (to which we refer in the sequel) can be defined as

$$I_s = \frac{S_n - E[S_n]}{n}. \tag{2.2}$$

The normalized index I_s converges in distribution as the number of leaves n grows to infinity. According to Rösler (1991), the limit X satisfies a (functional) fixed-point equation of the following type

$$X = UX + (1-U)X^* + 2U \log U + 2(1-U)\log(1-U) + 1 \tag{2.3}$$

where X, X^*, U are independent random variables, X and X^* are identically distributed, U is uniformly distributed over the inverval $(0, 1)$, and the equality holds for distributions. Using equation (2.3), the variance of the limiting distribution can be computed in an exact way

$$\sigma^2 = 7 - \frac{2\pi^2}{3}.$$

2.2.2 Number of cherries

McKenzie and Steel (2000) considered a simple and easily computed statistic for evaluating tree shape: the number of cherries of the tree. A *cherry* is a pair of leaves that are adjacent to a common ancestor node. The authors analysed the distribution of this statistic under the Yule model. They obtained exact

formulae for the mean and variance of the number of cherries, and showed that this distribution is asymptotically normal as the number of leaves grows to infinity. More specifically, if we denote by Ch_n the number of cherries in a tree of size n, their results can be summarized as follows

$$E[Ch_n] = \frac{n}{3}$$

and

$$\mathrm{Var}[Ch_n] = \frac{2n}{45}\,.$$

Using an argument based on extended Polya urns, McKenzie and Steel obtained that

$$\frac{Ch_n - n/3}{\sqrt{2n/45}} \to \mathcal{N}(0,1)\,.$$

2.2.3 The number of subtrees of fixed size

Sackin's statistic and the number of cherries are two distinct aspects of the same distribution: the number of leaves under a randomly chosen node. Let Z_n denote this number. On the one hand, Z_n is connected to Sackin index by the fact that

$$Z_n = \tilde{N}_J,$$

where J is a uniform random variable over that subset $\{1,\ldots,n-1\}$ (recall that $\tilde{N}_j$ is the number of leaves that descend from the ancestor j). Given the tree structure T, one actually has

$$E[Z_n \mid T] = \frac{1}{n-1}\sum_{j=1}^{n-1}\tilde{N}_j = \frac{S_n}{n-1}\,.$$

On the other hand, the empirical frequency of cherries in a Yule tree $f_n(2) = Ch_n/(n-1)$ is an unbiased estimator of the probability of the event $(Z_n = 2)$. The distribution of Z_n has been described by Blum and François (2005b) using results from coalescent theory. This distribution can be find in the Proposition 2 from the chapter 5 of the present manuscript.

Theorem 1 *(Blum and François, 2005b) Let $n \geq 2$ and Z_n be the number of individuals in a uniformly chosen random clade of a Yule tree with n leaves. We have*

$$p_n(z) = \mathrm{P}(Z_n = z) = \frac{n}{(n-1)} \frac{2}{z(z+1)}, \quad z = 2, \ldots, n-1,$$

and

$$p_n(n) = \mathrm{P}(Z_n = n) = \frac{1}{n-1}.$$

By the above remarks, we can check that $(n-1)Z_n$ has the same average value that Sackin's statistics (both are equal to the average complexity of the quicksort algorithm). In addition, we find that the frequencies of subtrees of size 2 is connected the number of cherries as follows

$$f_n(2) = \frac{Ch_n}{n-1} = \frac{1}{n-1} \sum_{j=1}^{n-1} \mathbb{1}_{(N_j=2)}$$

Taking expectation, this leads to

$$E[f_n(2)] = E[\frac{1}{n-1} \sum_{j=1}^{n-1} \mathbb{1}_{(N_j=2)}] = \mathrm{P}(N_J = 2) = \frac{n}{3(n-1)}, \tag{2.4}$$

and we can recover the fact that $E[Ch_n] = n/3$.

Normality of frequencies of subset sizes For a Yule tree T with n leaves, let $f_n(z)$ denote the frequencies of subtrees of size z in the tree ($z \geq 2$). Equation (2.4) tells that the frequency $f_n(2)$ is an unbiased estimator of the probability $p_n(2)$. The same argument applies to proving that $f_n(z)$ is an unbiased estimator of the probability $p_n(z)$ for all $z \geq 2$. In addition, we obtain that the limiting distribution of $f_n(z)$ is Gaussian as n goes to infinity.

Theorem 2 *Let $z \geq 2$. The empirical probabilities $f_n(z)$ have variances of order $1/n$*

$$\mathrm{Var}[f_n(z)] = \frac{\sigma^2 n}{(n-1)^2} \sim \frac{\sigma^2}{n}.$$

In addition, the following convergence in distribution holds

$$\sqrt{n}\,(f_n(z) - p_n(z)) \to \mathcal{N}(0, \sigma^2(z)), \quad \text{as } n \to \infty$$

where, for all $z \geq 2$,

$$\sigma^2(z) = \frac{2(z-1)(4z^2-3z-4)}{z(2z+1)(2z-1)(z+1)^2}.$$

Proof. Let $X_n^z = (n-1)f_n(z)$ denote the number of families of size z. The case $z = 2$ is a direct consequence of McKenzie and Steel's results (2000) because $X_n^2 = Ch_n$ is equal to the number of cherries in a coalescent or Yule tree. In addition, $\sigma^2(2)$ is equal to $2/45$. Let us give a sketch of proof for the general result. For $z \geq 3$, the random variable X_n^z can be involved into a quicksort-like recurrence equation (see Hwang and Neininger, 2002)

$$X_n = X_J + X_{n-J}^* + t_n \tag{2.5}$$

where J is uniformly taken from the set $\{1, \ldots, n-1\}$ and the *toll* function t_n is equal to

$$t_n = \delta_{n,z}$$

where $\delta_{n,z}$ denotes the Kronecker symbol. In this equation, X and X^* are independent copies which correspond to the values obtained for the left and right subtrees after a split at the root of the tree. The formal expression of the variance of X_n^z was computed using equation (2.5) and elementary programming in MAPLE. For $n \geq 2z+1$, we obtained that $\mathrm{Var}[X_n^z] = \sigma^2(z)n$ with $\sigma^2(z)$ given by the Theorem. The final result follows from Hwang and Neininger (2002). ■

Comments. A rather direct proof of Theorem 2 can be found in (Devroye, 1991). Devroye states the result for binary search trees. To obtain the above result, it must be modified according to the one-to-one correspondence between binary search trees and Yule trees. A binary search tree with $(n-1)$ nodes corresponds identically to a Yule tree with n leaves (see Aldous, 2001). Another proof for the mean and variance formulae was given by Rosenberg (2004) by purely

combinatorial techniques.

For all $z \geq 2$ and n sufficiently large, five percent level confidence intervals for frequencies are given by

$$-1.96\sigma(z)\frac{\sqrt{n}}{n-1} < f_n(z) < +1.96\sigma(z)\frac{\sqrt{n}}{n-1}$$

The accuracy of the approximation depends on z. For $z \leq 10$, simulation evidences show that Gaussian distributions provide good fit to the empirical distributions of $f_n(z)$ as soon as $n \geq 30$.

2.2.4 New indices

To conclude this section, we propose a new statistic based on the comparison of the empirical and theoretical distributions of the number of subtrees under the Yule model. This index is defined as a weighted ℓ_1 distance

$$D = \sum_{z=2}^{n} z\,|f_n(z) - p_n(z)|$$

where the sum runs over the all possible subset sizes under an arbitrary node.

In spirit, the metric D is comparable to the Sackin's index, giving importance to the apparition of abnormally large number of leaves under the nodes close to the root. However it has the advantage of providing a one-sided test, whereas Sackin index provides a two-sided test.

2.3 Distribution of indices

In this section, we estimate the quantiles of the distributions of the normalized Sackin's index and the distance index D for samples of various sizes.

Experimental design In order to estimate the quantiles of the distributions $10,000$ Yule trees were simulated. All simulations were performed under the object-oriented R language, which provide facilities for manipulating tree data

(R Development Core Team, 2003). Empirical cumulative distribution functions were computed and displayed in Figures 2.1 and 2.2. Approximate values of the empirical quantiles can easily be determined from these graphical representations.

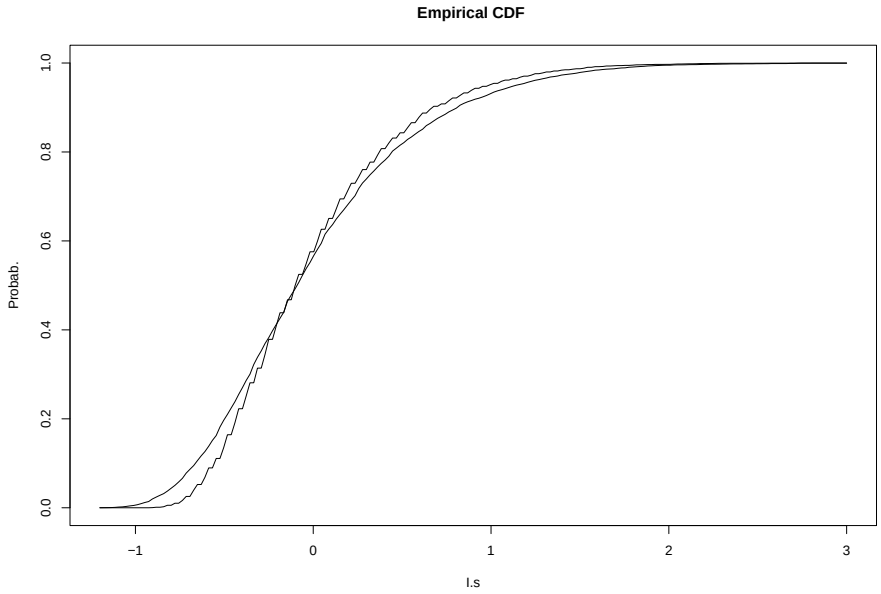

Figure 2.1: *Empirical cumulative distribution functions of the normalized Sackin's index under the Yule model for random trees of sizes* $n = 30$ *and* $n = 100$ *(smooth line). The cdfs were computed from 10,000 trees*

Normalized Sackin Normalized values of the Sackin index have been considered earlier by Kirkpatrick and Slatkin (1993). However, these authors proposed the value 1.96 for statistical significance at the five percent level (Gaussian approximation). We propose corrections that account for the fact that the distribution of the normalized Sackin's index is non Gaussian. For $n = 30$ taxa, the five percent rejection area can be described as

$$I_s < -0.72 \quad \text{or} \quad I_s > 1.24$$

For $n = 100$, the five percent rejection area is slightly different

$$I_s < -0.87 \quad \text{or} \quad I_s > 1.43.$$

The limiting distribution was computed numerically according to equation 5.3 (using a method developed by Tan and Hadjicostas (1995)), and good agreement with the empirical distribution was noticed for $n = 100$.

Distance D Regarding the distance D, no normalization is available theoretically. Empirical quantiles for samples of size $n = 30, 100, 200$ can be deduced from the Figure 2.2. The test is significant at the five percent level if $D > 5.10$ for $n = 30$, $D > 8.04$ for $n = 100$ and $D > 9.65$ for $n = 200$. Simulation results show that D should be corrected to $\tilde{D} = D - 2.27\log(n) + 3.62$ (almost perfect log-linear fit). The five percent rejection area is $\tilde{D} > 1.25$ for large n.

2.4 Statistical power

2.4.1 Biased speciation model

A basic issue regarding the power of these different statistics is which one is the most sensitive to a departure from the Yule model. This question is also relevant to coalescent models and population genetics where departure from neutrality is an important step in detecting the evolutionary pressures acting on a sample of genes.

The power of statistics depends on the alternative hypothesis and there are many possible choices that may produce imbalanced trees. For instance, Pinelis introduced a family of models that emcompass the traditional hypotheses about tree-biased speciation (Pinelis, 2003). Kirkpatrick and Slatkin (1993) evaluated the power of five tree shape statistics encompassing the Sackin and Colless indices at three tree sizes (10, 20 and 40 species). They generated imbalanced phylogenies by making the instantaneous rates of speciation of every pair of sister lineages differ by a constant factor. They concluded that the Sackin and Colless indices performed well (except for the smallest trees).

Agapow and Purvis considered other processes of non random speciation in which the rate of speciation in a lineage evolves inpendently of the rates in other lineages and is directed toward greater biological relevance (see Agapow

and Purvis, 2002). Among 8 studied statistics, they found that the Sackin and Colless indices performed well for models of trait evolution. These indices were a little less powerful when applied to a model of age-dependent rates.

Biased speciation model In this study, we used an alternative model of biased speciation which is similar to the one used by Kirkpatrick and Slatkin. Assume that the speciation rate of a specific lineage is equal to r ($0 \leq r \leq 1$). When a species with speciation rate r splits, one of its descendent species is given the rate pr and the other is given the speciation rate $(1-p)r$ where p is fixed for the entire tree. These rates are effective until the daughter species themselves speciate. Values of p close to 0 or 1 yield very imbalanced trees while values around 0.5 lead to over-balanced phylogenies.

We simulated this model for different numbers of species $n = 30, 100, 200$ and different values of p. The most interesting values are around $p = 0.12 - 0.15$ where it may sometimes be difficult to detect the imbalance visually. Type two errors β were calculed from $10,000$ independent Monte Carlo repetitions and the type one error α was fixed at the level of 5 percent.

Results The results are reported in Tables 2.1, 2.2, and 2.3. The performance of the statistics $f_n(z)$ to detect imbalance were weak for small ($n = 30$) phylogenies. They were sightly better for larger trees $n = 100 - 200$. Among subtrees, counting the number of cherries appeared to be the most efficient way of detecting departure from the Yule model. Overall, $f_n(z)$ and Ch_n showed very low power for all z, and we would not recommend their use for testing imbalance.

Sackin statistics I_s and the D distance were very powerful as concerned high desequilibrium, ie $p = 0.05$ for $n = 30$, and $p \leq 0.1$ for $n = 100, 200$. In this situation, I_s was sightly more powerful than D. When imbalance is less evident $p = 0.125 - 15$, the performances of both indices decreased. In this situation, we typically obtained 86 % of errors for D while this ratio was equal to 92 % for I_s. The decrease of power was thus slower for D. The last rows of Tables 2.1, 2.2 and 2.3 show that D was unable to detect over-balanced trees while I_s performed rather well in this regard (due to the two-sided test).

$n = 30$ p	D	I_s	$f_n(2)$	$f_n(3)$	$f_n(4)$	$f_n(5)$	$f_n(6)$	$f_n(7)$	$f_n(8)$
$p = 0.05$	0.02	0.005	0.79	0.80	0.95	1	–	–	–
$p = 0.1$	0.42	0.36	0.95	0.96	0.98	1	–	–	–
$p = 0.125$	0.69	0.72	0.96	0.96	0.99	1	–	–	–
$p = 0.15$	0.86	0.92	0.97	0.96	0.98	1	–	–	–
$p = 0.25$	0.99	0.98	0.97	0.95	0.98	0.99	0.99	0.96	0.99
$p = 0.4$	0.99	0.35	0.93	0.94	0.97	1	0.99	0.96	0.99
$p = 0.5$	1	0.11	0.92	0.93	0.97	1	1	0.95	0.98

Table 2.1: *Type two error β for the alternative hypothesis H_1 of biased speciation with parameters $n = 30$, and $p = 0.05 - 0.5$ computed from 10,000 repetitions. The type one error is $\alpha = 0.05$.*

$n = 100$ p	D	I_s	$f_n(2)$	$f_n(3)$	$f_n(4)$	$f_n(5)$	$f_n(6)$	$f_n(7)$	$f_n(8)$	$f_n(9)$	$f_n(10)$
$p = 0.05$	0	0	0.21	0.96	0.99	0.99	0.98	0.97	0.90	–	–
$p = 0.1$	0.01	0.00	0.55	0.88	0.99	0.98	0.97	0.94	0.94	–	–
$p = 0.125$	0.24	0.27	0.78	0.90	0.98	0.99	0.98	0.94	0.93	0.92	–
$p = 0.15$	0.77	0.90	0.91	0.92	0.98	0.99	0.98	0.96	0.94	0.96	–
$p = 0.25$	0.99	0.98	0.89	0.95	0.97	0.97	0.97	0.94	–	–	–
$p = 0.4$	0.99	0	0.63	0.92	0.96	0.96	0.96	0.96	0.92	–	–
$p = 0.5$	1	0	0.57	0.91	0.95	0.96	0.95	0.95	0.92	–	–

Table 2.2: *Type two error β for the alternative hypothesis H_1 of biased speciation with parameters $n = 100$, and $p = 0.05 - 0.5$ computed from 10,000 repetitions. The type one error is $\alpha = 0.05$.*

$n = 200$ p	D	I_s	$f_n(2)$	$f_n(3)$	$f_n(4)$	$f_n(5)$	$f_n(6)$	$f_n(7)$	$f_n(8)$	$f_n(9)$	$f_n(10)$
$p = 0.05$	0.02	0.01	0.001	0.75	0.99	0.90	0.87	0.88	0.92	0.92	0.96
$p = 0.1$	0.42	0.36	0.49	0.77	0.88	0.97	0.99	0.96	0.93	0.84	0.82
$p = 0.125$	0.69	0.71	0.83	0.92	0.91	0.94	0.98	0.98	0.97	0.92	0.90
$p = 0.15$	0.86	0.92	0.93	0.95	0.94	0.94	0.98	0.98	0.98	0.97	0.95
$p = 0.25$	1	0.92	0.81	0.93	0.93	0.93	0.97	0.97	0.97	0.95	0.95
$p = 0.4$	1	0.0	0.32	0.86	0.92	0.92	0.95	0.96	0.96	0.96	0.95
$p = 0.5$	1	0.0	0.24	0.85	0.91	0.92	0.95	0.96	0.96	0.96	0.95

Table 2.3: *Type two error β for the alternative hypothesis H_1 of biased speciation with parameters $n = 200$, and $p = 0.05 - 0.5$ computed from 10,000 repetitions. The type one error is $\alpha = 0.05$.*

2.5 Example

This section analyzes a dataset taken from the literature (Yusim et al., 2001). Several authors attempted to infer historical features of the acquired immune deficiency syndrome (AIDS) using human immunodeficiency virus type 1 (HIV-1) sequences. There are three distinctive form of HIV-1 (M,O,N). Group M contains the viruses which cause the global HIV pandemic and appear to have arisen in Central Africa during the last 100 years (Korber et al., 2000). Vidal et al. (2000) investigated the genetic diversity of HIV-1 group M in this region by obtaining viral gene sequences in 1997 from 197 infected individuals living in the Democratic Republic of Congo. Yusim et al. (2001) used a maximum likelihood approach to estimate a phylogeny for this large data set, and it is this phylogeny that we use here. The phylogeny is availabe from the R package *ape* (Paradis et al., 2004).

Korber et al. (2000) estimated the time since the most recent common ancestor of the HIV-1 thanks to this tree, and this was compared with a coalescent approach by Yusim et al. (2001). These works were based on the assumption of a molecular clock, or a constant rate of evolution among each lineage. Rambaut et al. (2001) noticed that it is unlikely that this HIV-1 data set has been evolving according to this hypothesis. Therefore, the goodness-of-fit of Yule or coalescent (essentially the same topological model) to this dataset has to be tested.

	$z=2$	$z=3$	$z=4$	$z=5$	$z=6$	$z=7$
theoretical	.335	.167	.100	.067	.047	.035
empirical	.312	.140	.114	.057	.057	.036

Table 2.4: *Theoretical and empirical distribution $f_n(z)$ of the size of subsets in the HIV-1 phylogeny group M. The number of sequences is $n = 192$, and z denotes the size of subsets.*

To assess the fit to a Yule model, we computed the normalized Sackin index, the empirical distribution of subset sizes, and the weighted ℓ_1 distance D. The Sackin's index was equal to $I_s = 0.82$ ($\mathrm{P}(I_s > 0.82) = 0.10$). The empirical distribution $f_n(z)$ of subset sizes is given in Table 2.4. The test provided by theorem 2

$$|f_n(z) - p_n(z)| > 1.96\sigma^2(z)\sqrt{n/(n-1)}$$

was non-significant for all $z \leq 10$. The distance D was equal to $D = 9.37$ ($\tilde{D} = 1.05$) and the p-value was equal $\mathrm{P}(\tilde{D} > 1.05) \approx 0.08$. We could not conclude to the rejection of the Yule or coalescent models.

To go further, we remark that imbalance might be detected at any level of the tree, and we considered cutting the branches of the tree which were far from the root. Doing so, we kept only the "old" internal branches that corresponded to the 30 oldest ancestors. Under the null hypothesis for the $n = 192$ sequences, the pruned tree should also be compatible with a Yule model. In this case, the Sackin's index was equal to $I_s = 1.21$ ($\mathrm{P}(I_s > 1.21) = 0.03$). The empirical distribution $f_n(z)$ of subset sizes is given in Table 2.5. The test provided by Theorem 2 was significant for $z = 2$ (cherries). The distance D was equal to $D = 5.17$ and the p-value was $\mathrm{P}(D > 5.17) \approx 0.04$. These results probably indicate a change in the evolutionary rate during the evolution which had more impact on cladogenesis during the early expansion of the virus.

	$z=2$	$z=3$	$z=4$	$z=5$	$z=6$	$z=7$
theoretical	.344	.172	.103	.068	.049	.036
empirical	.241	.172	.103	.068	.068	.068

Table 2.5: *Theoretical and empirical distribution $f_{30}(z)$ of the size of subsets in the HIV-1 phylogeny group M. The tree was pruned to keep the 30 oldest internal nodes (the top of the tree).*

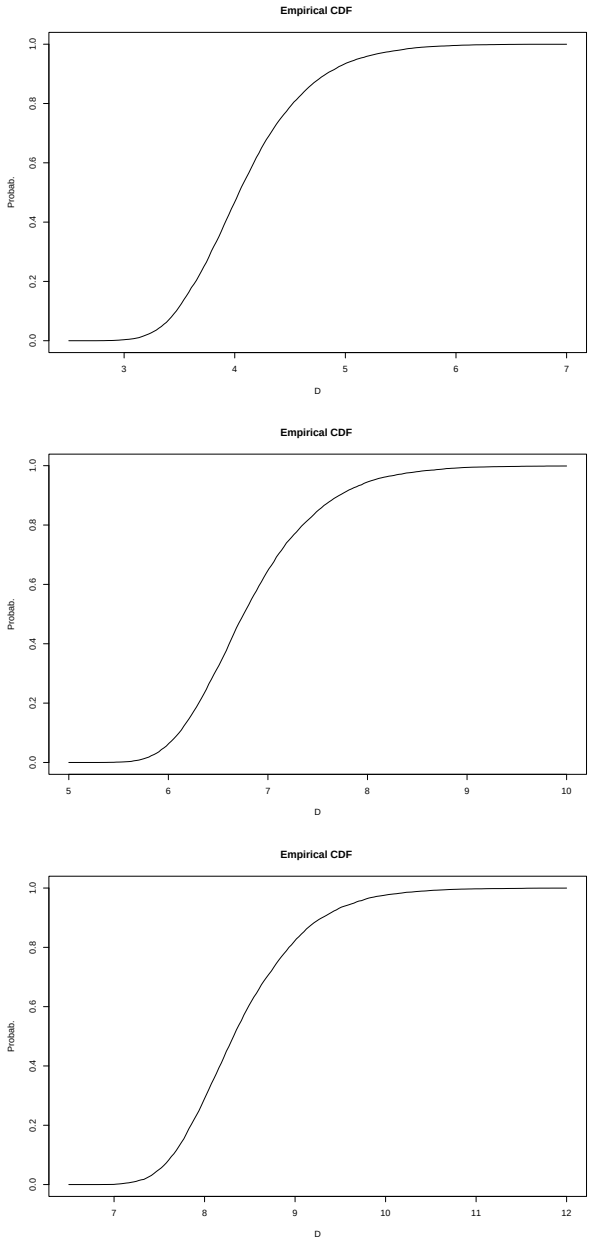

Figure 2.2: *Empirical cumulative distribution functions of the D statistic under the Yule model for random trees of sizes $n = 30$, $n = 100$, and $n = 200$.*

Chapter 3

The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance

Abstract

Two statistics called the Sackin's and Colless' indices are frequently used for capturing the balance of phylogenetic trees. In this article, these statistics are studied under the Yule and uniform model of phylogenetic trees. Asymptotics for the mean, variance, covariance of these indices are obtained, as well as their limiting joint distribution for large phylogenies. Under the Yule model the limiting distribution arises as a solution of a functional fixed point equation. Under the uniform model the limiting distribution is the Airy distribution. The cornerstone of this study is that the probabilistic models for phylogenetic trees are strongly related to the random permutation and the Catalan models for binary search trees.

3.1 Introduction

Phylogenetic trees (PT) represent the shared history of extant species. The idea of using trees to model evolution dates to Darwin (1859, see his diagram p.117). In a (rooted) PT, there is a common ancestral species called the root and each branching represents the time of divergence. A PT is usually reconstructed using data from n different species (or taxa) which are located at the leaves. The tree has $n-1$ internal nodes that correspond to the ancestors of the sample. There are two distinct features of rooted PT's. First is the branching structure or topology of the tree. Second is the branch lengths which indicate the time separating major evolutionary events. The shape of such trees carries useful information about the history of diversification rates among species by reflecting the footprint left by evolutionary processes.

Biologists have widely investigated the way by which the shape of PT's can be measured (Kirkpatrick and Slatkin, 1993). Mooers and Heard (1997) wrote an exhaustive review about tree balance in systematic biology, and Aldous (2001) gave a introduction in a more mathematical setting. How these measures are related to macroevolution processes has been studied by Rogers (1994) and Agapow and Purvis (2002) relying upon heavy computer simulations. So far several statistics have been introduced to measure the shape of PT's (see Agapow and Purvis, 2002, for eight of them). Among these statistics, the most widespread are the Sackin's and the Colless' indices. Sackin's index (Sackin, 1972, Shao and Sokal, 1990) computes the sum of depths for all leaves in the tree. Colless' index (Colless, 1982) inspects the internal nodes, partitioning the leaves that descend from them into groups of sizes L and R, and computes the sum of absolute values $|L-R|$ for all ancestors.

The probabilistic distribution of the Sackin's and Colless' statistics have been investigated for various models of biologically plausible random trees. Two random models of PT are often considered in the literature. The most famous is the Yule model (Yule, 1924). The Yule model is a branching process with constant speciation rate where the number of extant species is specified. The assumption of constant speciation rate may be weakened by assuming that the diversification rate could vary in time but is the same for all species at any time. This assumption

does not modify the distribution of PT shape. An alternative model considered by biologists is called the uniform model. It assumes that all PT's are equally likely. This model is biologically motivated as it arises from a large family of Galton-Watson processes conditioned by the total size of the trees (see Aldous, 1991a). In addition, McKenzie and Steel (2001) have shown that when speciation events are constrained to occur before a time τ after their previous speciation event, then the resulting process converges to the uniform model as τ tends to zero. In both models, Rogers (1994) studied the joint distribution of the Sackin's and the Colless' statistics using numerical computations. He concluded that these statistics were strongly correlated in large PT's. The limiting distribution of the Colless' statistic was also conjectured to be non Gaussian (Rogers, 1993).

This article describes the mean, the (co)variance and the limiting joint distribution of the Sackin's and Colless' indices for large PT's under the Yule and uniform models. Because this study is mainly concerned with the topology of PT's, branch lengths can be ignored. A PT is then a cycle-free connected graph, with vertices of degree one (the leaves), two (the root), or three (all ancestors except the root). Leaves are usually labeled whereas ancestors are not. This simplified model of phylogeny without branch lengths is sometimes called a cladogram (see Aldous, 1995). Our proofs use the connection to recent results in theoretical computer science, as well as the correspondence between PT's and random binary search trees (BST). This approach extends results by Blum and François (2005a) who showed that the Sackin's index has the same limit distribution as the number of comparisons used by the quicksort algorithm (Hoare, 1961). More specifically, we deal with the Yule and uniform models separately. For the Yule model, our analysis relies on the recursive structure of the tree, and makes use of the fixed point method (see for instance Hwang and Neininger, 2002). This method was introduced in the probabilistic analysis of algorithms by Rösler (1991). In the uniform model, the results are based on the connection between uniform trees and Bernoulli excursions (Takacs, 1991). A large family of statistics similar to the Sackin's and Colless' indices have been studied by Fill and Kapur (2004) under the Catalan model for BST's.

In Section 2, we shall present our main results. Section 3 explains how probabilistic models for PT's are related to probabilistic models for BST's. Section 4

is dedicated to the Yule model while Section 5 deals with the uniform model.

3.2 The Sackin and the Colless for large phylogenetic trees

Consider a PT with n leaves. Sackin's statistic adds the number of internal nodes between each leaf and the root of the tree to form the following index

$$S_n = \sum_{i=1}^{n} d_i \,,$$

where the sum runs over the n leaves of the tree and d_i is the number of ancestors crossed in the path from i to the root (including the root). Colless' statistic inspects the internal nodes, partitioning the leaves that descend from them into groups of sizes L_j and R_j and computes

$$C_n = \sum_{j=1}^{n-1} |L_j - R_j| \,,$$

where the sum runs over the internal nodes, and L_j (resp. R_j) corresponds to the number of leaves in the left (resp. right) subtree under node j.

Denote $\mathcal{M}_2$ the space of all bivariate, centered probability measures with finite second moments on $\mathbb{R}$, and $\mathcal{L}(X)$ an element of $\mathcal{M}_2$. We have the following result.

Theorem 3 *Assume the Yule model of PT. Consider the map $\mathcal{T} : \mathcal{M}_2 \to \mathcal{M}_2$ such that for all $\nu \in \mathcal{M}_2$ we have*

$$\mathcal{T}(\nu) = \mathcal{L}\left(\begin{bmatrix} U & 0 \\ 0 & U \end{bmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} + \begin{bmatrix} 1-U & 0 \\ 0 & 1-U \end{bmatrix} \begin{pmatrix} X_1' \\ X_2' \end{pmatrix} + \begin{pmatrix} b_S \\ b_C \end{pmatrix} \right),$$

with

$$\begin{pmatrix} b_S \\ b_C \end{pmatrix} = \begin{pmatrix} 2U\log U + 2(1-U)\log(1-U) + 1 \\ U\log U + (1-U)\log(1-U) + 1 - 2\min(U, 1-U) \end{pmatrix}$$

where (X_1, X_2), (X_1', X_2') *and* U *are independent random variables such that* $\mathcal{L}(X_1, X_2) = \mathcal{L}(X_1', X_2') = \nu$ *and* U *is uniform over the interval* $(0, 1)$. *Then, we have*

$$\left(\frac{S_n - \mathbb{E}[S_n]}{n}, \frac{C_n - \mathbb{E}[C_n]}{n} \right) \xrightarrow{d} (S, C), \quad n \to \infty$$

where the convergence holds in distribution, and the limiting probability distribution is the unique fixed-point of the map $\mathcal{T}$.

Comments. The convergence in Theorem 3 will actually be proved for a stronger topology than convergence in distribution. As can be seen from section 4, it indeed holds for the Wasserstein-Mallows d_2-metric (Rachev and Rüschendorf, 1995) which warrants the existence and convergence of the second moments.

Comments. This result extends the fact that the normalized Sackin index

$$\bar{S}_n = \frac{S_n - \mathbb{E}[S_n]}{n} \tag{3.1}$$

converges in distribution to the same limit as the number of comparisons in the quicksort algorithm. According to Rösler (1991), the limit S satisfies a (functional) fixed-point equation of the following type

$$S \stackrel{d}{=} US + (1-U)S' + 2U \log U + 2(1-U)\log(1-U) + 1 \tag{3.2}$$

where S, S', U are independent random variables, S and S' are identically distributed, U is uniformly distributed over the inverval $(0, 1)$, and the equality holds for distributions. Regarding the Colless' index, the functional fixed point equation becomes

$$C \stackrel{d}{=} UC + (1-U)C' + U \log U + (1-U)\log(1-U) + 1 - 2\min(U, 1-U)\,. \tag{3.3}$$

This is a well-known result in systematic biology that the expectation of S_n is

of order $2n \log n$. More precisely, Kirkpatrick and Slatkin (1993) showed that

$$\mathbb{E}[S_n] = 2n \sum_{j=2}^{n} \frac{1}{j}$$

and

$$\mathbb{E}[S_n] = 2n \ln n + (2\gamma - 2)n + o(n)\,,$$

where γ is the Euler constant. Using the connection to the quicksort algorithm, the variance of the limiting distribution can be obtained according to Knuth (1973) as

$$\mathrm{Var}[S_n] \sim \left(7 - 2\frac{\pi^2}{3}\right) n^2, \quad n \to \infty\,.$$

These results can be extended to the case of the Colless' index as follows, with the remark that S_n and C_n are strongly correlated for large PT's.

Theorem 4 *Assume the Yule model of PT. Then, we have*

$$\mathbb{E}[C_n] = n \log n + (\gamma - 1 - \log 2)n + o(n) \tag{3.4}$$

$$\mathrm{Var}[C_n] \sim (3 - \frac{\pi^2}{6} - \log 2)n^2 \tag{3.5}$$

$$\mathrm{Cor}[S_n, C_n] \sim \frac{27 - 2\pi^2 - 6\log 2}{\sqrt{2(18 - \pi^2 - 6\log 2)(21 - 2\pi^2)}} \approx 0.98 \tag{3.6}$$

as n goes to infinity.

Regarding the uniform model of PT, mathematical results have deserved less attention than for the Yule model. After an appropriate rescaling we prove the convergence of both S_n and C_n toward the same marginal probability distribution, and we identify this distribution as $\sqrt{8}$ times the integral of the standard Brownian excursion $e(t)$

$$\omega = \int_0^1 e(t)\, dt\,.$$

The distribution of random variable $\mathcal{A} = \sqrt{8}\omega$ is known as the Airy distribution. A formula for the moments of $\mathcal{A}$ is given by Flajolet and Louchard (2001). In particular, we have

$$\mathbb{E}[\mathcal{A}] = \sqrt{\pi}\,,$$

and

$$Var[\mathcal{A}] = \frac{10 - 3\pi}{3}.$$

Theorem 5 *Assume the uniform model of PT. Then, we have*

$$\left(\frac{S_n}{n^{3/2}}, \frac{C_n}{n^{3/2}}\right) \xrightarrow{d} (\mathcal{A}, \mathcal{A}), \quad n \to \infty \tag{3.7}$$

where the convergence holds in distribution.

Comments.

$$\frac{S_n}{n^{3/2}} \xrightarrow{d} \mathcal{A}$$

immediately. This result was actually established by Takacs (1991) using the method of moments. In addition we find that

$$\mathbb{E}[S_n] \sim \sqrt{\pi} n^{3/2}$$

and

$$\mathrm{Var}[S_n] \sim (\frac{10}{3} - \pi) n^3 .$$

The moments of C_n follow from the next Theorem.

Theorem 6 *Assume the uniform model of PT. Then, we have*

$$\mathbb{E}[C_n] \sim \sqrt{\pi} n^{3/2},$$

and

$$\mathrm{Var}[C_n] \sim \frac{10 - 3\pi}{3} n^3$$

as n goes to infinity. In addition, the variables S_n and C_n are asymptotically correlated

$$\mathrm{Cor}[S_n, C_n] \sim 1,$$

and we have for any $k, \ell \geq 0$

$$\mathbb{E}[C_n^k \, S_n^\ell] \sim n^{3(k+\ell)/2} \mathbb{E}[\mathcal{A}^{k+\ell}]$$

as n *goes to infinity.*

3.3 Phylogenetic and binary search trees

Trees are often encountered in theoretical computer science as data structures connected with *divide and conquer* algorithms. In this section, we explain how binary search trees can be mapped onto phylogenetic trees univoquely, and how probabilistic models for BST's are transported on probabilistic models for PT's.

Mapping binary search trees A binary tree can be defined recursively. It is either empty or it is a node (the root) with left and right subtrees. A binary search tree is a binary tree where labels are associated to the vertices. These labels are constrained: the label of a vertex is greater or equal than all labels contained in the left subtree and lower or equal than all labels contained in the right subtree (Sedgewick and Flajolet, Section 5.5, 1996). The transformation that maps BST's into PT's is explained by Aldous (1995). Given a BST with $n-1$ vertices, the structure is modified as follows. Vertices in the BST become ancestors in the PT. Do do so, two leaves are connected to each vertex of degree one, and one leaf is connected to each vertex of degree two. The root has a special status. If the degree is 0, 1 or 2, then 2, 1 or 0 leaves are added. The labels of leaves are chosen arbitrarily amongst the $n!$ possible orders.

Two obtained PT's are equivalent if their left and right subtrees can be swapped recursively (see Figure 3.1). The set of PT's is the set of equivalence classes for this equivalence relation. Figure 3.2 gives a graphical representation of these transformations.

A PT may therefore arise from the construction of 2^{n-1} equivalent modified BST's. Because there are $\mathcal{C}_{n-1}$ BST's with $(n-1)$ vertices, we obtain $\mathcal{C}_{n-1} n!/2^{n-1}$

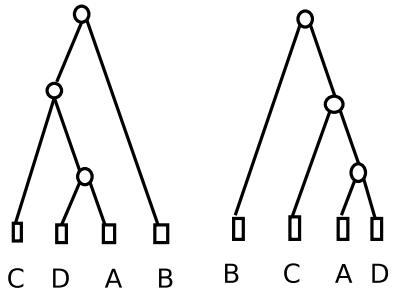

Figure 3.1: Two graphical representations of the same PT. They are identical by swapping left and righ subtrees.

possible PT's. This number coincides with the total number of PT's which equals $(2n-3)!!$.

Probabilistic models The mapping described in the above paragraph also transfers probabilistic models for BST's to probabilistic models for PT's. For instance, there will be an equivalence between the random permutation model for BST's (Mahmoud, 1992) and the Yule model for PT's. Probabilistic models for BST's (with $n-1$ vertices) can be described as a general class of models called *branching Markov processes*. A definition of branching Markov processes can be found in (Aldous, 1995). We recall this definition here. Let $\hat{q}_{n-1}$ be a symmetric probability distribution on $\{0,\ldots,n-2\}$

$$\hat{q}_{n-1}(i) = \hat{q}_{n-1}(n-2-i), \quad i = 0,\ldots,n-2\,.$$

In the branching Markov process, the size of the left subtree is chosen according to the probability distribution $\hat{q}_{n-1}$. This procedure is repeated recursively in subtrees assuming local independence. The probability distribution $\hat{q}_{n-1}$ is called the splitting distribution. In the same way, probability distributions on PT's

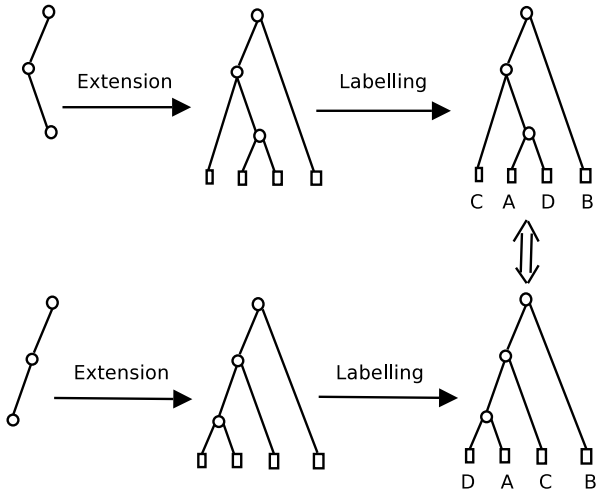

Figure 3.2: The transformation of 2 binary search trees to the same PT. The extension consists of connecting two leaves to vertices with outdegree 0 and one leaf to vertices with outdegree 1. The two obtained trees represent the same PT.

with n leaves can be associated with splitting probability distributions q_n on $\{1, \ldots, n-1\}$. At each step, the labels of the left subtree, of size i, are sampled uniformly amongst all the $\begin{pmatrix} n \\ i \end{pmatrix}$ possible labels. At the end of the construction, distinctions left-right are merely suppressed for building the PT.

Lemma 1 *Assume that $\hat{T}_{n-1}$ is a BST sampled according to a branching Markov process with splitting probability $\hat{q}_{n-1}$. Denote by Φ_n the transformation which consists of extending a BST with $n-1$ vertices into a PT with n leaves. Then $T_n = \Phi_n(\hat{T}_{n-1})$ is a PT sampled according to a branching Markov process with splitting probability q_n such that*

$$q_n(i) = \hat{q}_{n-1}(i-1), \quad i = 1, \ldots, n-1\,.$$

Proof. This is a consequence of the basic properties of Φ_n. If a BST has i vertices in its left subtree, the resulting PT has $i+1$ leaves in one of the two

subtrees of the root. The symmetry property of $\hat{q}_n$ insures that all members of the same equivalence class have the same probability of occurrence. ■

Lemma 1 has the interesting consequence that well-studied models of BST's can be shifted into models on PT's. The Yule and uniform models on PT's fall then into special cases of branching Markov processes.

On one hand the random permutation model for BST's with $n-1$ vertices is a branching Markov process with splitting probability

$$\hat{q}_{n-1}(i) = \frac{1}{n-1}, \quad i = 0, \ldots, n-2\,.$$

This model is mapped by Φ_n into the Yule model for PT's with n leaves with splitting probability

$$q_n(i) = \frac{1}{n-1}, \quad i = 1, \ldots, n-1\,.$$

The splitting distribution for Yule trees was found by Harding (1971). Note that the same splitting property also holds for coalescent tree topologies (Kingman, 1982).

On the other hand, the Catalan model for binary BST with $n-1$ vertices assumes that all $\mathcal{C}_{n-1}$ binary trees have the same probability of occurrence. The number of trees with left subtree of size is i is equal to $\mathcal{C}_i\mathcal{C}_{n-2-i}$. The Catalan model is a branching Markov process where the splitting distribution is given by

$$\hat{q}_{n-1}(i) = \frac{\mathcal{C}_i\mathcal{C}_{n-2-i}}{\mathcal{C}_{n-1}}, \quad i = 0, \ldots, n-2\,.$$

The transformation Φ_n maps the Catalan model into the uniform model for PT with n leaves. The splitting distribution for PT is then

$$q_n(i) = \frac{1}{2}\left(\begin{array}{c} n \\ i \end{array}\right)\frac{(2i-3)!!(2(n-i)-3)!!}{(2n-3)!!}, \quad i = 1, \ldots, n-1\,.$$

This formula can also be found in (Aldous, 1995).

Lemma 2 *Let q_n be a splitting distribution on $\{1, \ldots, n-1\}$. Let h be a function of pairs of integral numbers. Denote by X_n an additive random variable defined recursively as*

$$X_n \stackrel{d}{=} X_{I_n} + X_{J_n} + h(I_n, J_n)$$

where the I_ℓ's are sampled under a branching Markov process of splitting distribution q_n and $I_n + J_n = n$. Define $\hat{X}_n$ by

$$\hat{X}_n \stackrel{d}{=} \hat{X}_{\hat{I}_n} + \hat{X}_{\hat{J}_n} + h(\hat{I}_n + 1, \hat{J}_n + 1)$$

where the $\hat{I}_\ell$'s are sampled under the branching Markov process of splitting distribution $\hat{q}_n$ with

$$\hat{q}_n(i) = q_{n+1}(i+1), \quad i = 0, \ldots, n-1$$

and $\hat{I}_n + \hat{J}_n = n - 1$. Then, we have

$$X_n \stackrel{d}{=} \hat{X}_{n-1}\,.$$

Proof. Note that $\hat{I}_n \stackrel{d}{=} I_{n+1} - 1$, i.e. the distribution of $\hat{I}_n$ is given by $\hat{q}_n$. Similarly, we have

$$X_n \stackrel{d}{=} X_{\hat{I}_{n-1}+1} + X_{\hat{J}_{n-1}+1} + h(\hat{I}_{n-1} + 1, \hat{J}_{n-1} + 1).$$

Setting $\hat{X}_{n-1} = X_n$, we prove the result. ■

Comments. This lemma states that for additive random variables built from a Markov branching PT of splitting distribution q_n there exist additive random variables $\hat{X}_n$ built from a Markov branching BST of distribution $\hat{q}_n$. In addition, X_n and $\hat{X}_{n-1}$ have the same distribution. This lemma can obviously be generalized to multivariate random variables. In the next sections, all random variables are studied in the context of BST's. Applying Lemma 2, the results can be shifted to PT's without difficulties.

3.4 Yule model

The Yule model is a branching Markov process for PT's with splitting probability

$$q_n(i) = \frac{1}{n-1}, \quad \text{for } i = 1 \dots (n-1)\,.$$

The Sackin's index S_n has been defined as a sum of depths over the leaves. It can also be expressed as a sum over the ancestors (Rogers, 1994)

$$S_n = \sum_{j=1}^{n-1} N_j$$

where N_j is the number of leaves descending from the ancestor j. Applying Lemma 2 we obtain that S_n has the same distribution as $\hat{S}_{n-1} + 2(n-1)$ where $\hat{S}_n$ is defined by

$$\hat{S}_n \stackrel{d}{=} \hat{S}_{I_n} + \hat{S}'_{J_n} + n - 1\,. \tag{3.8}$$

where I_n is distributed uniformly over $\{0, \dots, n-1\}$ and $J_n = n-1-I_n$. The recursion satisfied by $\hat{S}_n$ is well-studied as it arises from the analysis of the quicksort algorithm or the internal path length of a BST under the random permutation model. Similarly, C_n has the same distribution as $\hat{C}_{n-1}$, where

$$\hat{C}_n \stackrel{d}{=} \hat{C}_{I_n} + \hat{C}'_{J_n} + |I_n - J_n|\,. \tag{3.9}$$

In order to describe the joint distribution of $(\hat{S}_n, \hat{C}_n)$ under the random permutation model, we shall follow the same lines of proof as Neininger (2002) who studied the joint convergence of the Wiener index and the internal path length of a BST.

Proof of Theorem 3 *Step 1 (Computing expectations).* Denote $\hat{c}_n = \mathbb{E}[\hat{C}_n]$ and $\hat{s}_n = \mathbb{E}[\hat{S}_n]$. We have

$$\hat{s}_n = 2n \log n + (2\gamma - 4)n + o(n)\,. \tag{3.10}$$

We rewrite equation (3.9) as

$$\hat{C}_n \stackrel{d}{=} \hat{C}_{I_n} + \hat{C}'_{J_n} + n - 1 - 2\min(I_n, J_n)\,. \tag{3.11}$$

Conditioning on I_n in the above equation, we find that

$$\hat{c}_n = (n - 1 - 2t_n) + \frac{2}{n}\sum_{k=0}^{n-1} \hat{c}_k$$

where

$$t_n = \mathbb{E}\left[\min(I_n, J_n)\right] = \begin{cases} \frac{n-2}{4} & \text{if } n \text{ is even} \\ \frac{(n-1)^2}{4n} & \text{if } n \text{ is odd.} \end{cases}$$

The equation satisfied by $\hat{c}_n$ is solved by computing the difference $n\hat{c}_n - (n-1)\hat{c}_n$ and iterating the resulting recurrence (Hwang and Neininger, Lemma 1 p. 1691). We obtain that

$$\hat{c}_n = (n - 1 - 2t_n) + 2(n+1)\sum_{k=1}^{n-1} \frac{k - 1 - 2t_k}{(k+1)(k+2)}\,.$$

An asymptotic expansion of the above expression leads to the following result

$$\hat{c}_n = n\log n + (\gamma - 1 - \log 2)n + o(n)\,. \tag{3.12}$$

Step 2. Limit distribution. Let us consider the rescaled quantities

$$\hat{X}_n = \begin{pmatrix} \frac{\hat{S}_n - \hat{s}_n}{n} \\ \frac{\hat{C}_n - \hat{c}_n}{n} \end{pmatrix},$$

and $\hat{X}'_n$ an independent copy of $\hat{X}_n$. From equations (3.8) and (3.11) we have

$$\hat{X}_n = A_1^{(n)}\hat{X}_{I_n} + A_2^{(n)}\hat{X}'_{J_n} + b^{(n)} \tag{3.13}$$

where

$$A_1^{(n)} = \frac{1}{n}\begin{pmatrix} I_n & 0 \\ 0 & I_n \end{pmatrix},$$

$$A_2^{(n)} = \frac{1}{n}\begin{pmatrix} J_n & 0 \\ 0 & J_n \end{pmatrix},$$

and

$$\begin{pmatrix} b_S^{(n)} \\ b_C^{(n)} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \hat{s}_{I_n} + \hat{s}_{J_n} - \hat{s}_n + n - 1 \\ \hat{c}_{I_n} + \hat{c}_{J_n} - \hat{c}_n + n - 1 - 2\min(I_n, J_n) \end{pmatrix}.$$

Since I_n/n converges in L^2 toward U a uniform variable over $(0,1)$, we obtain that

$$A_1^{(n)} \xrightarrow{L^2} A_1^* = \begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix},$$

and

$$A_2^{(n)} \xrightarrow{L^2} A_2^* = \begin{pmatrix} 1-U & 0 \\ 0 & 1-U \end{pmatrix}.$$

Using the asymptotic expansion of $\hat{c}_n$ given by equation (3.12) and the asymptotic expansion of $\hat{s}_n$ given by (3.10), we find that

$$b_S^{(n)} = \frac{1}{n}\left(2I_n \log\left(\frac{I_n}{n}\right) + 2J_n \log\left(\frac{J_n}{n}\right) + n\right) + o(1)$$

and

$$b_C^{(n)} = \frac{1}{n}\left(I_n \log\left(\frac{I_n}{n}\right) + J_n \log\left(\frac{J_n}{n}\right) + n - 2\min(I_n, J_n)\right) + o(1).$$

Thus, we obtain that

$$\begin{pmatrix} b_S^{(n)} \\ b_C^{(n)} \end{pmatrix} \xrightarrow{L^2} b^* := \begin{pmatrix} 2U\log U + 2(1-U)\log(1-U) + 1 \\ U\log U + (1-U)\log(1-U) + 1 - 2\min(U, 1-U) \end{pmatrix}.$$

Assuming that $\hat{X}_n$ converges in distribution, the limiting distribution $\mathcal{L}(\hat{X})$ must satisfy the following condition

$$\hat{X} \stackrel{d}{=} A_1^*\hat{X} + A_2^*\hat{X}' + b^* \tag{3.14}$$

where $\hat{X}$, $\hat{X}'$,and (A_1^*, A_2^*, b^*) are independent and $\hat{X}' \stackrel{d}{=} \hat{X}$.

The multivariate contraction theorem (Neininger, 2001) states that there is

a unique probability distribution $\mathcal{L}(\hat{X})$ satisfying (3.14) in $\mathcal{M}_2$. Moreover it states that the distribution of $\hat{X}_n$ converges toward the distribution of $\hat{X}$ in the Wasserstein-Mallow d_2-metric. The convergence in this metric is the same as the convergence in distribution and the convergence of second moments. Neininger's Theorem can be applied provided that the four following conditions hold

(i) $(A_1^{(n)}, A_2^{(n)}, b^{(n)}) \stackrel{L^2}{\to} (A_1^*, A_2^*, b^*),\ n \to \infty,$

(ii) $\mathbb{E}\left[\|(A_1^*)^t A_1^*\|_{op}\right] + \mathbb{E}\left[\|(A_2^*)^t A_2^*\|_{op}\right] < 1,$

(iii) $\mathbb{E}\left[1_{\{I_n \le \ell\} \cup \{I_n = n\}} \|(A_1^*)^t A_1^*\|_{op}\right] \to 0,\ \forall \ell \in \mathbb{N},\ n \to \infty,$

(iv) $\mathbb{E}\left[1_{\{J_n \le \ell\} \cup \{J_n = n\}} \|(A_2^*)^t A_2^*\|_{op}\right] \to 0,\ \forall \ell \in \mathbb{N},\ n \to \infty,$

where $\|A\|_{op} = sup_{\|x\|=1}\|Ax\|$ is the operator norm of A. For symmetric matrices, as here, this equals the spectral radius.

(i) has already been proved. (ii) holds because

$$\mathbb{E}\left[\left\|(A_1^*)^t A_1^*\right\|_{op}\right] + \mathbb{E}\left[\left\|(A_2^*)^t A_2^*\right\|_{op}\right] = \mathbb{E}[U^2 + (1-U)^2] = \frac{2}{3} < 1\,.$$

(iii) and (iv) are obvious because $\|(A_r^*)^t A_r^*\|_{op} \le 1$, for $r = 1, 2$, and

$$P(\{I_n \le \ell\} \cup \{I_n = n\}) = P(\{J_n \le \ell\} \cup \{J_n = n\}) \le \frac{\ell + 1}{n} \to 0$$

for all $\ell \in \mathbb{N}$ and $n \to \infty$. ■

Proof of Theorem 4 According to Theorem 3, equation (3.14) has an unique solution, so we can consider (S, C), (S', C'), two independent copies with $\mathcal{L}(S, C) = \mathcal{L}(S', C')$ being the fixed point of (3.14) and U uniform over $(0, 1)$. By definition, we have

$$\mathcal{L}\begin{pmatrix} S \\ C \end{pmatrix} \stackrel{d}{=} \mathcal{T}\mathcal{L}\begin{pmatrix} S \\ C \end{pmatrix}$$

Using the fact that all random variables (except U) are centered, we find that

$$\begin{aligned}\mathbb{E}[C^2] &= \mathbb{E}[U^2C^2] + \mathbb{E}[(1-U)^2C'^2] \\ &\quad + \mathbb{E}[(U\log U + (1-U)\log(1-U) + 1 - 2\min(U,1-U))^2]\,.\end{aligned}$$

Thus, we have

$$\mathrm{Var}[C^2] = \mathbb{E}[C^2] = (3 - \frac{\pi^2}{6} - \log 2)\,.$$

In the same way, we find that

$$\begin{aligned}\mathbb{E}[SC] &= \mathbb{E}[U^2SC] + \mathbb{E}[(1-U)^2S'C'] \\ &\quad + \mathbb{E}[(2U\log U + 2(1-U)\log(1-U) + 1) \\ &\quad * (U\log U + (1-U)\log(1-U) + 1 - 2\min(U,1-U))]\,.\end{aligned}$$

This leads to

$$\mathrm{Cov}(S,C) = \mathbb{E}[SC] = \frac{9}{2} - \frac{\pi^2}{3} - \log 2\,.$$

Using that $\mathbb{E}[S^2] = 7 - 2\pi^2/3$ (Rösler, 1991), we find that

$$\mathrm{Cor}(S,C) = \frac{27 - 2\pi^2 - 6\log 2}{\sqrt{2(18 - \pi^2 - 6\log 2)(21 - 2\pi^2)}}\,.$$

Theorem 3 holds in the Wasserstein-Mallows d_2-metric which implies the convergence of second moments. This leads to

$$\mathrm{Var}[S_n] \sim \mathrm{Var}[S]n^2, \qquad \mathrm{Var}[C_n] \sim \mathrm{Var}[C]n^2,$$

and

$$\mathrm{Cov}(S_n,C_n) \sim \mathrm{Cov}(S,C)n^2, \qquad \mathrm{Cor}(S_n,C_n) \sim \mathrm{Cor}(S,C)\,.$$

■

3.5 Uniform model

For a given n, the uniform model assumes that all PT's with n leaves are equally likely. Again we use the fact that the Sackin's and the Colless' indices for a PT

with n leaves drawn according to the uniform model have the same probability distribution as $\hat{S}_{n-1}+2(n-1)$ and $\hat{C}_{n-1}$ which are respectively defined by equation (3.8) and (3.9). Under the Catalan model for BST, I_n is distributed according to $\hat{q}_n$, where

$$\hat{q}_n(i) = \frac{\mathcal{C}_i \mathcal{C}_{n-1-i}}{\mathcal{C}_n}, \quad i = 0, \ldots, n-1 \,.$$

Conditional on I_n, $(\hat{S}_{J_n}, \hat{C}_{I_n})$ is independent of $(\hat{S}'_{J_n}, \hat{C}'_{J_n})$.

Clearly, $\hat{S}_n$ has the same distribution as the internal path length of a BST under the Catalan model, and $\hat{C}_n$ has the same distribution as the random variable $\sum_{j=1}^{n-1} |\hat{L}_j - \hat{R}_j|$ where the sum is over the $n-1$ vertices of a BST drawn under the Catalan model. Note that $\hat{C}_n$ can be rewritten as

$$\sum_{j=1}^{n-1} (\hat{N}_j - 1) - 2\min(\hat{L}_j, \hat{R}_j)$$

where $\hat{N}_j$ is the number of vertices of the subtree rooted at j (including j), and $\hat{L}_j$ ($\hat{R}_j$) is the number of vertices of the left (right) subtree. Then we have

$$\hat{C}_n = \hat{S}_n - 2\sum_{j=1}^{n-1} \min(\hat{L}_j, \hat{R}_j) \,.$$

It is well known (Takacs, 1991) that $\hat{S}_n/n^{3/2}$ converges in distribution to the Airy distribution. The proof relies on the one-to-one correspondence between binary trees and Bernoulli excursions. Takacs (1991) computed the moments of $\hat{S}_n/n^{3/2}$ and their limiting values to establish convergence from the method of moments. The goal of this section is to prove the convergence of

$$\mathbb{E}\left[\frac{\sum_{j=1}^{n-1} \min(\hat{L}_j, \hat{R}_j)}{n^{3/2}}\right] \to 0, \quad n \to 0 \,.$$

This implies that $(\hat{C}_n - \hat{S}_n)/n^{3/2}$ converges in probability to 0. By the Slutsky's Theorem, this argument completes the proof of Theorem 5.

Lemma 3 *Let $n \geq 2$, and consider a BST with n vertices under the Catalan*

model. Denote j_0 the root of the tree. We have

$$\mathbb{E}[\min(\hat{L}_{j_0}, \hat{R}_{j_0})] \leq K\sqrt{n}$$

for some constant K.

Proof. The Stirling's formula yields to a well known asymptotic expansion for the Catalan number $\mathcal{C}_n$

$$\mathcal{C}_n = \frac{4^n}{\sqrt{\pi n^3}} \left(1 + O(\frac{1}{n})\right) . \tag{3.15}$$

The expectation of $\min(\hat{L}_{j_0}, \hat{R}_{j_0})$ is given by

$$\mathbb{E}[\min(\hat{L}_{j_0}, \hat{R}_{j_0})] = \sum_{k=1}^{\lfloor n/2 \rfloor - 1} 2k \frac{\mathcal{C}_k \mathcal{C}_{n-1-k}}{\mathcal{C}_n} + \mathbb{1}_{\{n \in 2\mathbb{N}+1\}} \frac{(n-1)\mathcal{C}^2_{(n-1)/2}}{2\mathcal{C}_n}$$

Using (3.15), we find that

$$\sum_{k=1}^{\lfloor n/2 \rfloor - 1} 2k \frac{\mathcal{C}_k \mathcal{C}_{n-1-k}}{\mathcal{C}_n} = \frac{4^{n-1}}{\pi \mathcal{C}_n (n-1)} \frac{1}{n-1} \sum_{k=1}^{\lfloor n/2 \rfloor - 1} 2 \frac{(1 + O(1/k) + O(1/n))}{\sqrt{k/(n-1)}(1 - k/(n-1))^{3/2}}$$

The sum in the right hand side of the above equality is a Riemann sum. Using that

$$\int_0^{1/2} \frac{1}{\sqrt{x}(1-x)^{3/2}} = 2 ,$$

we have

$$\sum_{k=1}^{\lfloor n/2 \rfloor - 1} 2k \frac{\mathcal{C}_k \mathcal{C}_{n-1-k}}{\mathcal{C}_n} = \frac{4^n}{\pi \mathcal{C}_n (n-1)} (1 + o(1)) .$$

Using (3.15) again leads to

$$\mathbb{E}[\min(\hat{L}_{j_0}, \hat{R}_{j_0})] \sim \sqrt{n/\pi} . \tag{3.16}$$

which concludes the proof of the result. ■

By conditioning on the sizes of the two subtrees of the root, and using induction on n, it follows that

$$\mathbb{E}\left[\frac{\sum_{j=1}^{n-1}\min(\hat{L}_j,\hat{R}_j)}{n^{3/2}}\right] \leq K\mathbb{E}\left[\frac{\sum_j\sqrt{\hat{N}_j}}{n^{3/2}}\right].$$

In the following, we prove that the right-hand side of the above inequality converges to 0 as n grows to ∞. A lemma which is interesting in its own gives the key argument. In the following, we use the standard convention that $\begin{pmatrix} 0 \\ 0 \end{pmatrix} = 1$.

Lemma 4 *Let $n \geq 1$. Consider a BST with n vertices sampled according to the Catalan model. Pick a vertex V at random amongst the n vertices. Denote $\hat{K}_n = \hat{N}_V$, we have*

$$\mathbb{P}(\hat{K}_n = k) = \frac{\mathcal{C}_k \begin{pmatrix} 2n-2k \\ n-k \end{pmatrix}}{n\mathcal{C}_n}, \qquad \textit{if } k = 1,\ldots,n\,.$$

Proof. The proof relies on combinatorial arguments. Let us denote $\nu_k(T)$ the number of subtrees with k vertices in the BST T with $|T| = n$ vertices. For $k = 1,\ldots,n$, we have

$$\begin{aligned}\mathbb{P}(\hat{K}_n = k) &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{P}(\hat{N}_j = k)\\ &= \frac{1}{n}\mathbb{E}[\sum_{j=1}^{n}\mathbb{1}_{\{\hat{N}_j=k\}}] = \frac{1}{n}\mathbb{E}[\nu_k(T)]\,.\end{aligned}$$

$\nu_k(T)$ satisfies the following linear recursion

$$\nu_k(T) = \delta_{|T|,k} + \sum_S \nu_k(S) \tag{3.17}$$

where δ denotes the Kronecker symbol and the sum is over the subtrees of the root of T. Let us denote by $\mathcal{B}$ the set of all BST's. We introduce the cumulative

generating function defined by

$$F_k(z) = \sum_{T \in \mathcal{B}} \nu_k(T) z^{|T|},$$

and

$$G_k(z) = \sum_{T \in \mathcal{B}} \delta_{|T|,k} z^{|T|} = \mathcal{C}_k z^k.$$

From the linear recurrence equation (3.17), Theorem 5.7 in (Sedgewick and Flajolet, 1996) establishes a relationship between cumulative generating functions F_k and G_k

$$F_k(z) = \frac{G_k(z)}{\sqrt{1-4z}}\,.$$

Using that

$$\frac{1}{\sqrt{1-4z}} = \sum_{i \geq 0} \binom{2i}{i} z^i,$$

we find that

$$F_k(z) = \sum_{i \geq k} \binom{2i-2k}{i-k} \mathcal{C}_k z^i\,.$$

Since the expectation of $\nu_k(z)$ is given by

$$\mathbb{E}[\nu_k(T)] = [z^n] F_k(z)/\mathcal{C}_n\,.$$

This completes the proof of the Lemma. ■

Corollary 1 *Let $n \geq 2$. Consider a PT with n leaves sampled from the uniform model. Let K_n denotes the number of leaves descending from a uniformly chosen random ancestor. We have*

$$\mathbb{P}(K_n = k) = \frac{\mathcal{C}_{k-1} \binom{2n-2k}{n-k}}{(n-1)\mathcal{C}_{n-1}}, \qquad \textit{for } k = 2, \ldots, n\,.$$

Proof. It is a direct consequence of Lemma 2. ■

Comments. As n grows to infinity, the distribution of $\hat{K}_n$ converges to

$$P(\hat{K}=k) = 4^{-k}\mathcal{C}_k \sim \frac{1}{\sqrt{\pi}k^{3/2}} \quad \text{for large } k.$$

The tail of the distribution of $\hat{K}$ has a power law of parameter 3/2. This can be compared to a similar result in the context of BST's (Martinez et al., 1998) under the random permutation model. In this case, $\hat{K}_n$ has power law distribution of parameter 2. Since 3/2 is lower than 2, large random subtrees are more likely in the Catalan model than in the random permutation model. It was an expected result since Catalan binary trees are known to be more unbalanced than BST's under the random permutation model (Sedgewick and Flajolet, section 5.6, 1996).

Comments. Actually, the limiting distribution in the preceding comment is equal to the size of the critical Galton-Watson process with a binomial Bi(2,1/2) offspring distribution (Aldous, Lemma 9, 1991b). This is the Galton-Watson process corresponding to binary trees.

Now we are ready to prove that $\mathbb{E}[\sum_j \sqrt{\hat{N}_j}]/n^{3/2}$ converges to 0 as n goes to ∞. Let $\alpha \in]0,1[$ and split the sum $\mathbb{E}[\sum_j \sqrt{\hat{N}_j}]$ in two parts

$$\mathbb{E}\left[\sum_j \sqrt{\hat{N}_j}\right] = \mathbb{E}\left[\sum_{j,\hat{N}_j<n^\alpha} \sqrt{\hat{N}_j} + \sum_{j,\hat{N}_j\geq n^\alpha} \sqrt{\hat{N}_j}\right]$$

Obviously, we have

$$\begin{aligned}\frac{1}{n^{3/2}}\mathbb{E}\left[\sum_{j,\hat{N}_j<n^\alpha} \sqrt{\hat{N}_j}\right] &\leq \frac{1}{n^{3/2}}\mathbb{E}\left[\sum_{j,\hat{N}_j<n^\alpha} \sqrt{n^\alpha}\right] \\ &\leq n^{\alpha/2-1/2} \to 0\end{aligned}$$

when n grows to ∞. For the second term, we have

$$\frac{1}{n^{3/2}}\mathbb{E}\left[\sum_{j,\hat{N}_j\geq n^\alpha} \sqrt{\hat{N}_j}\right] \leq \frac{1}{n}\mathbb{E}\left[\sum_{j,\hat{N}_j\geq n^\alpha} 1\right].$$

The right hand-side of the inequality is equal to $P(\hat{K}_n \geq n^\alpha)$. Applying Lemma 4, we find that

$$\frac{1}{n^{3/2}}\mathbb{E}\left[\sum_{j,\hat{N}_j \geq n^\alpha} 1\right] \sim \kappa\, n^{-\alpha/2}$$

for some constant κ. This expression converges to 0 when n grows to ∞. This completes the proof of Theorem 5.

Comments. Fill and Kapur (2004) established more precise results concerning $\mathbb{E}[\sum_j \sqrt{\hat{N}_j}]$. They proved that

$$\mathbb{E}[\sum_j \sqrt{\hat{N}_j}] \sim \frac{1}{\sqrt{\pi}} n \log n\,. \tag{3.18}$$

Their results rely on Hadamard products. Note that our proof uses elementary arguments, and is instructive in its own as concerns the shape of Catalan trees. Besides, equation (3.18) follows easily from Lemma 4, by direct estimation of the sum by an integral.

Proof of Theorem 6. In the following, we prove the convergence of the mixed moments of $\hat{S}_n$ and $\hat{C}_n$ which states that for $k, l \geq 0$, we have

$$\mathbb{E}[\hat{S}_n^k \hat{C}_n^\ell] \sim n^{3(k+\ell)/2}\mathbb{E}[\mathcal{A}^{k+\ell}]\,. \tag{3.19}$$

The argument is similar to the argument given by Janson (Remark 3.5, 2003) to establish the convergence of the mixed moments of the internal path length and the Wiener index. Since the convergence in distribution has been established in Theorem 5, the above equation is equivalent to uniform integrability of $n^{-3(k+\ell)/2}\hat{S}_n^k \hat{C}_n^\ell$ for $n \geq 1$ and any fixed k, ℓ. Since $\hat{C}_n \leq \hat{S}_n$, the result follows from the fact that $n^{-3k/2}(\hat{S}_n)^k$ is uniformly integrable for every fixed k. This is true because Takacs (1991) obtained the convergence of the moments of $\hat{S}_n$. ■

Chapter 4

Fertility inheritance unbalances gene genealogies

Abstract

We develop here a method to infer a cultural trait, namely fertility inheritance, from genetic data in human populations. This method is based on the reconstruction of the gene genealogy of a sample of sequences from a given population and on the computation of the degree of imbalance of this genealogy. We show indeed that this level of imbalance increases with the level of fertility inheritance, and that other phenomena like hidden population structure are unlikely to generate a signal of imbalance in the genealogy that would be confounded with fertility inheritance. Finally, we apply our method to mtDNA samples from 34 human populations. It shows that fertility inheritance is much more frequent in hunter-gatherer populations than in agriculturist populations.

4.1 Introduction

Some non-genetic departures from neutral evolution have been reported in human populations (see e.g. Heyer et al., 2005). Here, we are interested in populations where the progeny size of an individual is positively correlated to his/her parent progeny size. In that case, individuals whose parents had many children are more likely to have numerous offspring in these populations (Cavalli-Sforza and Feldman, 1981). This fertility inheritance has been explained in some cases by a cultural comportment transmission (Williams and Williams, 1974). In humans, founding a big family can be seen as a cultural behavior related to family structure and it has been shown that cultural traits related to the family are mainly transmitted by the parents (Guglielmino et al., 1995). Cultural inheritance is even not confined to humans: It has been reported in species of matrilineal whales (Whitehead, 1998).

Some effects of fertility inheritance on genetic diversity have been already studied. For instance, Nei and Murata (1966) showed that it strongly reduces the effective population size. Austerlitz and Heyer (1998) showed that it explains the high frequencies of some genetic diseases in several populations and Austerlitz and Heyer (2000) showed that it has some consequences on the mapping of loci involved in genetic disorders since it increases the level of association between a disease locus and closely-linked marker genes.

Until now, this intergeneration correlation of offspring size has been detected thanks to genealogical databases (Austerlitz and Heyer, 1998). Recently, Austerlitz et al. (2003) developed an indirect way to detect fertility inheritance thanks to genetic and demographic data. Basically, their method uses haplotypic data to estimate jointly the age of a given mutant allele and the growth rate of the number of carriers of this allele since its appearance. When the estimated growth rate is higher than the known population growth rate for several independent mutations, it is likely that a demographic phenomenon like fertility transmission is occuring in the population. This was shown in two populations, one of Bulgarian gypsies (Vlax) and one of Quebec (Saguenay-Lac Saint Jean).

In our study, we aim to detect this fertility correlation with genetic data only

using coalescent methods. It is indeed well-known that population demography has an effect on the coalescent tree shape (Nordborg, 2001). For instance, genealogies in an exponentially growing population will tend to be more starlike (Slatkin and Hudson, 1991; Di Rienzo and Wilson, 1991). Similarly, a departure from the neutral coalescent is found when fertility is inherited. In particular, Sibert et al. (2002) showed that fertility transmission makes coalescence times decrease. This reduction is higher in the branches near the MRCA, giving a starlike shape to the coalescent tree. However, since a similar shape is also expected in an expanding population, inferring only the lengths of the branches of the tree is not enough to infer fertility inheritance. However, Sibert et al. (2002) found a specific signature of fertility transmission: It increases the level of imbalance of coalescent trees. The imbalance of a tree is defined as the average level of imbalance of its nodes, knowing that a given node is completely balanced if it splits the sample in two subsamples of same sizes and it is all the more imbalanced that the sizes of the two subsamples differ.

While weakly studied until now in the coalescence theory, the balance of phylogenetic trees has been widely studied in systematic biology to test hypothesis about the macroevolutionary processes (Moers and Heard, 1997). Different measures computing the balance of the whole tree (Kirkpatrick and Slatkin,1993) as well as measures for a single node have been proposed. In this paper, we show how one of these whole tree balance measure, the mean I' (Agapow and Purvis, 2002) can be used as a method for detecting fertility correlation in a population from a sample of DNA sequences. Our method consists in reconstructing the coalescent tree of these sequences using classical phylogenetical methods, and to compute the mean I' value of this tree and its level of significance. A significant value of I' is considered as the result of fertility transmission. To assess the validity of our method, we performed first a power study, based on repeated simulations, to show the impact of the level of fertility transmission on the mean I' of the coalescent trees in a given population. Second, we tested the robustness of our method under different demographic scenarios. Third, we investigated also the effect of the phylogenetic method used. Finally, we applied our test on 34 mtDNA samples from different human populations and studied whether fertility transmission is more frequently detected in traditional hunter-gatherer populations (HGPs) or in

agriculturist populations.

4.2 Tree Balance Measures

We focus here on the imbalance of the trees, ignoring branch lengths. Labelled trees without any explicit time scale are called cladograms (Aldous, 1995). Most of the statistics capturing tree imbalance (Kirkpatrick and Slatkin, 1993) assume that trees are fully resolved. This assumption will often not be fullfilled: when several sampled individuals carry exactly the same sequence, as in many cases (e.g. mtDNA in HGPs, Excoffier and Schneider, 1999), the gene genealogies cannot be reconstructed entirely. However, the fact that a branch leads to several individuals rather than one is informative in terms of imbalance. Fusco and Cronk's (1995) method modified by Purvis et al. (2002) has been devised to handle incompletely resolved trees. In this method, only the subtrees with more than three tips (i.e. with more than one topology for a given tree size) are considered. For each node giving rise to such kind of subtree, it computes:

$$I = \frac{B - m}{M - m}$$

where B is the size of the largest daughter clade, $m = \lceil \frac{n}{2} \rceil$ (respectively $M = n-1$) is the minimum (respectively maximum) value for B. In order to devise a statistic whose expected value is independent of n, Purvis et al. (2002) proposed the following modification:

$$\begin{aligned} \text{if } n \text{ is odd: } I^{'} &= I \\ \text{if } n \text{ is even: } I^{'} &= \frac{n-1}{n} I. \end{aligned}$$

To compute a summary statistic for the whole tree, Agapow and Purvis (2002) (16) considered the mean of $I^{'}$ for all nodes where the phylogeny is resolved. If mean $I^{'}$ is higher than 0.5, the tree is more unbalanced than a neutral coalescent tree. Since it is normalized, trees of different sizes can be compared using mean $I^{'}$. It is also worth noticing that all nodes contribute equally to mean $I^{'}$, and since there are more nodes near the tips of the tree, this statistic is mostly influenced

by these nodes near the tips (Agapow and Purvis, 2002). To assess whether mean I' is significantly higher than .5, the expected value for a neutral coalescent tree, we adopted the same randomization procedure as Agapow and Purvis (2002). For each tree, 5000 randomizations were performed. One randomization consists of replacing, for all the nodes of the tree, I' by $1 - I'$ with probability 1/2. The P-value for the neutral coalescent hypothesis is the fraction of means computed on randomized trees that are above or equal to the observed mean. As stressed by Agapow and Purvis (2002), the randomization test can be applied to incompletely resolved trees. All statistical computations were performed with the free software R, using in particular the APE package (Paradis,2004), which is dedicated to phylogenetic tree analysis.

4.3 Simulation study

4.3.1 Model for the coalescent with fertility correlation

We used a simulation approach that allowed us to compare Kingman's (1982) coalescent, which corresponds to the classical Wright-Fisher model without fertility transmission, to cases where fertility is inherited. While several models have studied fertility correlation (e.g. Caballero, 1994; Campbell, 1999; Austerlitz and Heyer, 1998), Sibert et al.'s model (2002) was the more straightforward to use here, since it is the only one to be a direct extension of the Wright-Fisher model.

This Wright-Fisher model describes the evolution of a haploid population of constant size N, where at each generation t, the parent of each individual is drawn at random with replacement among the individuals from generation $t-1$. Each parent has the same probability to be drawn. On the other hand, in Sibert et al.'s model (2002), the probability of a given individual i to be drawn as a parent depend on its own parent's progeny size, in other words its sibship size, denoted s_i. The probability for i to be chosen as a parent is set as $\frac{s_i^\alpha}{\sum_{j=1}^{n} s_j^\alpha}$, where α is the intensity of fertility inheritance ($\alpha \geq 0$). A value of $\alpha = 0$ corresponds to the classical Wright-Fisher model without fertility inheritance.

4.3.2 Impact of fertility transmission on tree imbalance

Populations of constant size $N = 5000$ were simulated according to the fertility inheritance model described above. The program is basically the simulation program developed by Sibert et al. (2002), only improved for a better use of memory. Setting α to a given value, we simulated the population for a high number of generations and sampled at random n individuals from the last generation. Since we stored all relevant population lineages, we were able to trace back the complete genealogy of the sampled individuals. We performed repeated simulations for 20 values of α ranging from 0 and 2 and computed in each case the mean and standard deviation of the observed mean I'. This allowed us also to determine the threshold value of α above which I' is expected to diverge significantly from 0.5 and so fertility transmission can be detected.

4.3.3 Robustness of the method

The simulations above assume an isolated population of constant size, where fertility transmission is constant over time. However, these assumptions are often violated in real cases. Therefore we performed simulations in various cases, namely expanding populations, populations where fertility transmission is only transitory, and structured populations. For the first two scenarios, we performed repeated simulations to compute the mean and sd of I' for the same range of values of α as above to determine whether the threshold value of α above which fertility transmission can be detected by I' is affected or not. For the population structure case, we perfomed simulations without fertility transmission to determine whether it could yield a spurious signal of fertility transmission.

Population expansion

We simulated two kinds of expanding populations. The first one is a population under continuous geometric growth (N(t+1)= k N(t)) whith growth rate $k = 1.01$. In the second one, the population size remains constant until 10 or 20 generations before present, where the growth is geometric (either $k = 1.2$ or $k = 1.4$).

Fertility transmission during a limited period of time

Assuming strong fertility inheritance during a high number of generations might be unrealistic. For instance, in the French Canadian population studied by Austerlitz and Heyer (1998), fertility inheritance was observed in the genealogical databases during approximately ten generations. To take this into account, we simulated populations that experienced fertility inheritance only during a period of time that started T generations before present and lasted during τ generations. T ranged from 0 to 60 generations and τ was either 5 or 10 generations.

Spatial structure

We may wonder if fertility inheritance is the only demographic or genetic process to produce imbalanced genealogies? Przeworsky et al. (1999) showed that low selection does not modify tree shape significantly. Barton (1998) as well as Fay and Wu (2000) suggested that selective sweeps can unbalance the coalescent at a locus close to the one under selection. The effect of various types of selection on tree imbalance could be investigated. However, we restricted ourselves here to the effect of geographical structure, since, like fertility transmission, it is a demographic process and it affects therefore not only a single locus but the whole genome. Coalescent trees with 100 individuals were simulated with the software SIMCOAL (Excoffier et al., 2000) under three models of spatial structure.

- A two islands model with equal and unequal effective migration rates. N_1 and N_2 denote the effective population sizes of the two populations, m_{ij} denotes the migration rate from population i to population j and n_i denotes the number of sampled individuals in population i.

- A model of population merging: 2 separated populations (of effective size N_1 and N_2) merged T generations ago and gave rise to an unique population.

- The range expansion model introduced by Excoffier (2004): it starts from a unique population (the mainland), of effective size $N_{mainland}$, from which an instantaneous expansion arose T generations ago. During this instantaneous expansion, all individuals from the continent colonize an infinite number of

islands. After the expansion, each island has an effective size of N_{island} and exhanged migrants with all the other islands, with the same migration rate (m).

4.3.4 The effect of tree reconstruction

Because gene genealogies are hardly known in practice but only reconstructed from DNA data sets, the effect of tree reconstruction on imbalance needs to be investigated. For this, we simulated coalescent trees with 100 individuals for several values of α. The fertility inheritance was only experienced during 10 generations beginning 20 generations ago. DNA sequences were generated along the coalescent trees using Seq-Gen (Rambaut and Grassly, 1997), according to the Hasegawa-Kishino-Yano (HKY85) substitution model, assuming equal base frequencies and a transition-transversion ratio of 4 (Hasegawa et al., 1985). The mutation rate was set either at $5 * 10^{-5}$/site/generation, as estimated for human mtDNA from pedigree studies (Parsons et al., 1997) or at $5 * 10^{-6}$/site/generation as estimated from the divergence time between humans and chimps (Tamura and Nei, 1993). The sequence length was assumed of 600bp. The coalescent trees of the simulated sequences were reconstructed either with UPGMA or a maximum-likelihood method. UPGMA trees were built with PHYLIP (Felsenstein, 1989) from a ML distance matrix estimated using the HKY85 model (the transition ratio was estimated). The maximum-likelihood reconstructions were performed with PHYML (Guindon and Gascuel, 2003) assuming BIONJ tree (Gascuel, 1997) as starting tree, using also the HKY85 model with an estimated transition ratio. ML trees were rooted thanks to a simulated outgroup. We performed 100 replicates for each value of α, and reconstructed the coalescent tree with both methods in each case. Then we counted for each method the number of cases in which a significant value of I' was obtained at the .1 level.

4.3.5 Experimental data

We applied our method to mitochondrial data sets from the database MOUSE, which is a compilation of mtDNA from hypervariable regions I and II of the D-

Table 4.1: The power of the mean I' test to detect tree imbalance at the .1 level for several values of the fertility inheritance parameter α

n \ α	0	.33	.66	1	1.33	1.66	2
20	.04	.06	.09	.19	.68	.75	.76
30	.14	.10	.12	.26	.74	.83	.90
50	.08	.15	.11	.45	.91	.97	.99
100	.04	.15	.08	.69	1	1	1

100 coalescent trees were simulated for each value of α and n.

loop. Samples of sequences from 345 populations are available. Since mtDNA is inherited maternally, mtDNA tree balance will be sensitive to fertility transmission from mother to daughter only. Samples containing less than 45 individuals were discarded. mtDNA samples from the region II of the D-loop were not included in the study since they are not available for most of the populations. The gene trees were built with PHYML (Gascuel and Guindon, 2003) and rooted with *Pan paniscus* sequences also available in MOUSE. Trees with less than 4 resolved nodes were discarded.

4.4 Results

4.4.1 Power of tests based on tree imbalance

The statistic mean I' reveals strong imbalance for $\alpha > 1$ (Figure 4.1). Mean I' increases linearly from $\alpha = 0$ to $\alpha = 2$ (the regression between the statistic and α in that range gives $R^2 = .52$). It suggests that more and more nodes become imbalanced as α increases. Therefore, the imbalance signal is not confined to the basal nodes only. However, the statistics did not deviate from the imbalance expected under the neutral model for $\alpha < 1$. The transition is surprisingly rapid between the two states. For $\alpha = 1$, we recall that the propensity to reproduce depends linearly on the sibship size. The linear dependence is therefore a transition state which splits the detectable fertility inheritance from the undetectable one.

In order to investigate the power of the tests, simulations were run for different values of n and α. Results are shown in Table 4.1. The type I error was fixed

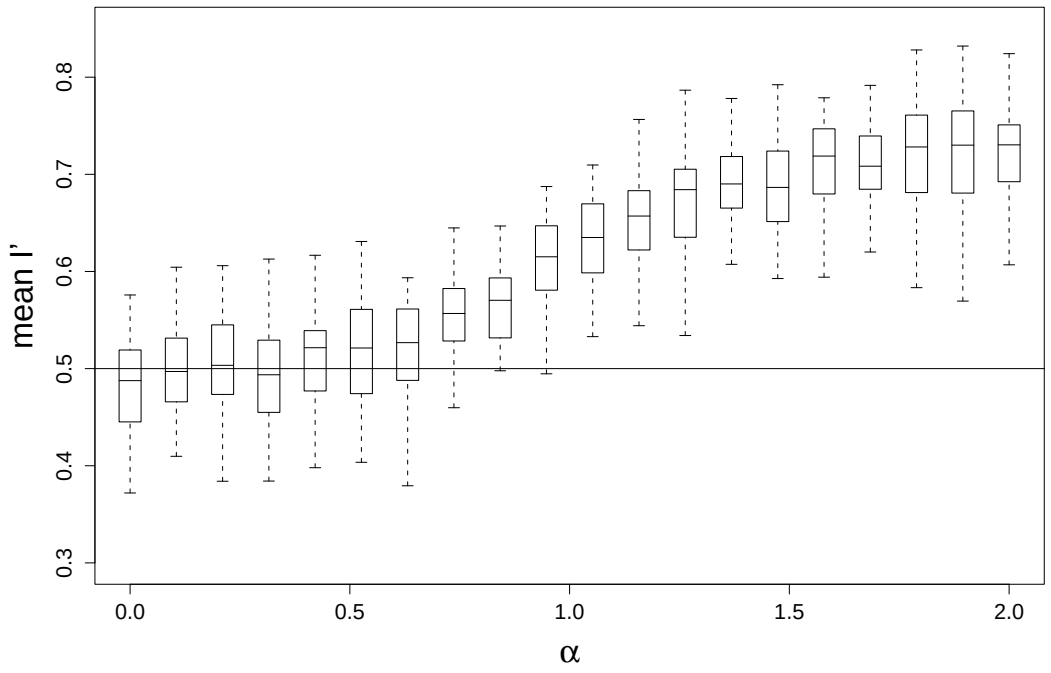

Figure 4.1: The effect of fertility inheritance on the balance measure mean I'. For each value of α, 100 coalescent trees with 100 individuals have been simulated. A box is delimited by the first and the third quartile. The horizontal line in each box represents the median and the whiskers extend to the most extreme data points. The horizontal line indicates .5 which is the expected value of mean I' under the neutral coalescent.

to 10%. As we noticed above, the test is unable to detect fertility inheritance for $\alpha < 1$, even when $n = 100$. For $\alpha \geq 1$, the test performs well even when the sample sizes are small (Power=.68 for $n = 20$ and $\alpha = 1.33$). For large sample sizes and $\alpha > 1$, the power of the test is close to 1.

4.4.2 Robustness of the method

Population expansion

The gene genealogies simulated in both models exhibited exactly the same imbalance pattern as in the constant population size model (results not shown).

Fertility inheritance in a short period of time

When fertility correlation occurs in the last generations, the pattern of imbalance remains identical to the one shown in Figure 1. In contrast, genealogy imbalance is detected in a population that experienced $\tau = 10$ generations of high fertility inheritance $T = 110$ generations ago. The transition between these two extreme scenarios is continuous. The threshold value of α above which fertility inheritance could be detected increased all the more that τ decreased and T increased (Figure 4.2). We may notice that strong fertility inheritance leaves an high fingerprint on genetic diversity even if it happens during five generations only.

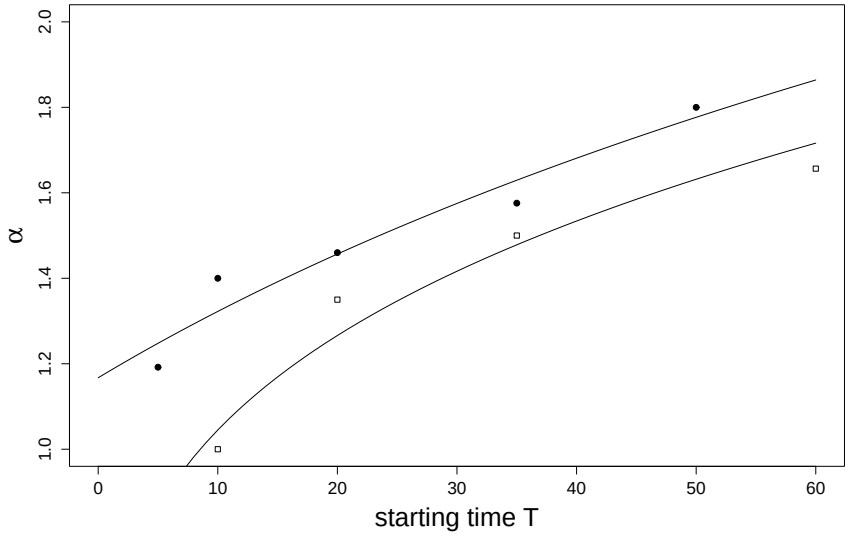

Figure 4.2: The effect of the starting time T of the fertility inheritance process on the α value above which it is possible to detect imbalance trees. The starting time is counted backward. The circles (respectively the squares) correspond to a model where fertility inheritance was experienced during 5 (respectively 10) generations. The solid lines have been obtained using linear regression.

The effect of tree reconstruction

The genetic diversity of the population for the both mutation rates is shown in Table 4.2. As also shown by Sibert et al. (2002) diversity decreases when fertility inheritance increases. This reduction is even more drastic for the low mutation

Table 4.2: The effect of fertility inheritance on heterozygosity

Substitution rate α	0	0.33	0.66	1	1.33	1.66	2
$\mu = 5 * 10^{-5}$/site/generation	.99	.99	.99	.99	.85	.65	.57
$\mu = 2.5 * 10^{-6}$/site/generation	.94	.93	.93	.93	.60	.15	.08

100 coalescent trees with 100 individuals were simulated for each value of α.

rate. If this rate was assumed, it is impossible to reconstruct gene genealogy for high fertility inheritance ($\alpha > 1$). In the following, only the larger mutation rate will be considered.

The power of imbalance tests on reconstructed genealogy is shown in Table 4.3. For $\alpha = 0$, 21% of the PHYML trees and 38% of the UPGMA trees are more unbalanced than predicted by the neutral coalescent. The expected number of rejected replicates is 10%. It means that reconstructed gene genealogies are slighty more unbalanced than expected under the neutral coalescent. The same bias has been reported in phylogeny where reconstruction methods provide biased estimate of tree shape (Huelsenbeck and Kirckpatrick, 1996). The UPGMA method introduce a substantial amount of imbalance and its use shall be proscribed. For α strictly bigger than 1, all the UPGMA topologies are more unbalanced than predicted by the neutral coalescent. For the same range of α (except $\alpha = 2$) most of the PHYML topologies are more unbalanced than predicted by the null hypothesis. For $\alpha = 2$ the PHYML seems to perform badly since only 54% of the reconstructed trees are significantly unbalanced. For $\alpha = 2$, the simulated DNA sequences are not polymorphic enough and this affects the quality of the tree reconstruction. When $\alpha = 2$, an individual with 3 brothers has a propensity to reproduce which is 16 times bigger that a single child. Such a big fertility inheritance is highly unlikely. In brief, the balance of the PHYML trees is quite close from the balance of the genealogies from which the DNA sequences have been generated, except for impropable high fertility inheritance.

Spatial structure

The effect of spatial structure on tree balance is shown in Table 4.4. In the two islands model, only unequal effective migration rates ($N_1 m_{12} \neq N_2 m_{21}$) can

Table 4.3: The power of the mean I' test to detect tree imbalance at the .1 level using reconstructed gene genealogies

Reconstruction method α	0	.33	.66	1	1.33	1.66	2
PHYML	.21	.18	.20	.44	.96	.79	.52
UPGMA	.38	.32	.32	.63	1	1	1

For each value of α, 100 coalescent trees with 100 individuals have been simulated. 100 sequences with 600bp have been drawn along the coalescent trees. The reconstruction methods have then been performed from the simulated sequences.

unbalance trees. In that case, the neutral coalescent hypothesis will be rejected at most 37% of the replicates. The sampling procedure which modifies the proportion of individuals coming from each population has no effect on tree balance. Scenarios of population fusion do not generate imbalance. For intermediate migration rates ($m = .01$), the range expansion model can unbalance coalescent trees at most 29% of the replicates. When migration rates are high, migration events are frequent and the islands look like a single panmictic population. In that case, genealogical trees have the same balance than expected under the neutral model. Although spatial structure may generate tree balance, it does not in the same proportion as fertility inheritance

Experimental data

The balance of the human mtDNA reconstructed trees are shown in Table 4.5 for HGPs and in Table 4.6 for other human populations. The number of resolved nodes is highly variable (mean= 12.66 and s.d.= 10.73) and positively correlated to the number of sequences available ($R^2 = .55$). There is no significant difference between the number of resolved nodes in the trees from the two different kind of populations ($p = .3$ for a one-sided Wilcoxon rank test). The heterozygosity is signifantly lower in HGPs ($p = .0034$ for a one sided Wilcoxon rank test) and correlated to the tree balance index ($p = .04$ for a Spearman rank test).

Fertility correlation appears much more frequent in HGPs. Indeed, only 7 out of 23 non HGPs showed a significant mean I' ($p \leq .05$), while it was the case for 8 out of 11 hunther-gatherers populations. The mean of the imbalance index was

Table 4.4: The power of mean I' test to detect tree imbalance at the .1 level under various spatial structure models

Spatial structure					Power of mean I′ test
2 islands					
N_1	N_2	$m_{12} = m_{21}$	n_1	n_2	
10000	10000	.001	100	0	.10
10000	10000	.001	50	50	.09
10000	10000	.01	100	0	.10
10000	10000	.01	50	50	.09
20000	5000	.001	50	50	.32
20000	5000	.01	100	0	.30
20000	5000	.01	0	100	.35
20000	5000	.01	50	50	.37
fusion					
N_1	N_2	T			
10000	10000	100			.10
10000	10000	1000			.12
20000	5000	100			.11
range expansion					
$N_{continent}$	N_{island}	$\sum N_{island}$	m	T	
1000	1000	100000	.01	1000	.28
1000	1000	100000	.1	1000	.12
1000	1000	100000	.01	100	.29

1000 coalescent trees with 100 tips were simulated for each parameter setting.

Table 4.5: mtDNA genealogy imbalance for hunter-gatherer populations

Populations	n	mean $I^{'}\star$	P value$\star$	$H\dagger$	resolved nodes$\star$
Aborigene Australian	50	.64	.24	1	8
Ainu	51	.80	.01	.95	8
Nuu − chah − nulth	63	.84	0	.95	7
Outer − island − yap	133	.60	.21	.93	8
Palau	139	.86	0	.93	21
Saami	177	.88	0	.90	8
Vanuatu	75	.79	.01	.98	12
Western new South Wales	63	.72	.02	.92	9
Yanomana	53	.75	.05	.93	8
Yap	202	.60	.13	.94	22
Yuendumu australian desert	51	.75	.01	.96	9

The genealogies were built from human mtDNA HV1 sequences.

0.75 (s.d.=0.10) in HGPs and 0.62 (s.d.=0.12) in non HGPs, the difference being significant ($p = 0.003$ for a one-sided Wilcoxon rank test).

$\star$ Mean $I^{'}$ values, their P values and the number of resolved nodes are obtained from reconstructed PHYML genealogies.

† H denotes the estimated heterozygosity.

Table 4.6: mtDNA genealogy imbalance for agriculturist populations

Populations	n	mean $I'\star$	P value$\star$	$H\dagger$	resolved nodes$\star$
Adygei	157	.64	.02	.98	22
Basque	106	.59	.29	.97	10
Bosnia	179	.44	.72	.98	8
Buryat	134	.62	.07	.98	22
Canarian	54	.71	.05	.98	8
Crete	186	.63	.07	.97	20
Croatian	64	.30	.88	.99	7
England	142	.66	.10	.98	13
Finns	50	.76	.03	.98	9
Fulbe	61	.71	.05	.97	11
Icelander	394	.65	0	.99	65
Indian Uttar Pradesh	73	.87	0	.99	7
Japanese	62	.58	.35	.99	8
Karelians	83	.73	.16	.96	5
Korean	378	.57	.14	.99	37
Lambadi	86	.57	.27	.99	12
Lobana	62	.67	.15	.98	11
Mandenka	120	.54	.31	.98	13

The genealogies were built from human mtDNA HV1 sequences.

4.5 Discussion

The simulation study shows that tree balance is a convenient way for detecting fertility inheritance. Different statistics have been proposed to capture tree balance (Kirkpatrick and Slatkin, 1993). The statistic that we used here (mean I', Agapow and Purvis, 2002) seems to be well adapted to gene trees in population genetics since it is designed to deal with not fully resolved trees. Moreover, we establish that tree balance is not affected by variation in population size and not too sensitive to geographical structure. Since the genealogical tree has to be built from genetic data, attention has to be paid to the quality of this reconstruction and the PHYML method appears to be the more efficient. Indeed the UPGMA method tends to unbalance trees, and its use would lead to inapropriate rejections of the neutral coalescent. It is encouraging to notice that the HIV1 tree of the Icelandic population is unbalanced (Icelandic HV2 tree is also unbalanced, result

Table 4.6: mtDNA genealogy imbalance for agriculturist populations

Populations	n	mean $I^{'}\star$	P value$\star$	$H\dagger$	resolved nodes$\star$
Mongols	103	.46	.65	.99	18
Ryukyuan	50	.74	.03	.99	6
Sardinian	69	.46	.64	.94	7
Serbian	64	.70	.08	.99	9
Uyghur	45	.68	.09	.99	8

The genealogies were built from human mtDNA HV1 sequences.

not shown) which is consistent with the detection of fertility intergenerational correlation from icelandic genealogical database (Helgason, 2003).

One striking result of the analysis of the data available in MOUSE is that mtDNA genealogies from HGPs are more unbalanced than the ones from non HGPs. This observation suggests that fertility inheritance is likely to be involved in these HGPs. This result is far from beeing intuitive since these populations do not control their fertility by contraception. The resulting number of children of an individual seems therefore more related to biological constraints than to cultural traits. However this concept of "natural fertility population" which refers to groups whose individuals have no preconceived target family size is controversial (see Campbell and Wood, 1986). We may sensibly object that tree balance is the consequence of diverse phenomenon which are more likely to occur in HGPs and that are not related to fertility inheritance.

The heterozygosity in HGPs is lower than the ones from non HGPs (See Results). This reduced genetic diversity may affect the reconstruction methods. The imbalance of trees would reflect bias in tree reconstruction methods only. Recent bottlenecks that have occured in these groups after the Neolithic transition has been proposed to explain their reduced genetic diversity (Excoffier and Schneider, 1999). Even if we assume that the bias introduced by the reconstruction methods are not too high, we don't have a guarantee that we have detected a demographic or cultural process. Since we have screened the HV1 gene only, we may have detect selection involving this gene. If this gene has a strong impact on any of the proximate determinants of fertility (Campbell an Wood, 1998) -the factors that have a direct impact on fertility- it may have unbalanced the HV1 genealogies.

Why this selection process would have been more pronounced in HGPs remains an open question.

It should be stressed that our method detects strong fertility inheritance only. Strong fertility inheritance means $\alpha > 1$ which corresponds to a fertility correlation greater than .2 (Sibert et al., 2002). The correlation values which are computed in demographic studies (Pearson's statistic) are not "real" intergenerational fertility correlation: the heterogeneity of the population (see Murphy, 2001) as well as a correlation between fertility and time periods (see Helgason et al., 2002) may increase this statistic. Truly speaking, our method detects very strong fertility inheritance only, i.e. a correlation indice greater than .2 once all the contributions to the Pearson' statistic except fertility inheritance has been removed. However if this high rate of fertility correlation has been reached during just a few generations, we have shown that it may still modify tree balance. To get an idea, the correlation indices are usually between 1.5 and 2 for modern western countries with a low fertility (Murphy, 2001). In the Quebecian populations studied by Auterlitz and Heyer (1998), it is equal to .161 and it is much higher (.34) if they restricted their study to the 18^{th} century. For HGPs, the only estimation of the correlation indice that we have found is provided by Draper and Hames (2000). They studied the !Kung from Botswana and found a correlation of .255 which raised to .447 if they restricted the analysis to males and fell to .076 if they restricted the analysis to females. Nontheless, it should be noted that all these correlation indices are not computed exactly in the same way. Austerlitz and Heyer (1998) computed the correlations of effective family size. By effective family size, they mean the number of offsprings who effectively reproduce per reproducing women. The other correlation index, called below the usual correlation, has been computed for all pairs parents-offsprings (including the offsprings who don't have any children). The former indice may be larger that the latter one. For instance, the usual correlation index of the Saguenay-Lac Saint Jean population is equal to .07 whereas the correlation of effective family size is equal to .161. Our method detect correlation of effective family size that is higher than .2 (Sibert et al., 2002). So it may detect usual correlation (the one that takes all the pairs parent-offspring into account) that is lower than .2. The data of Draper and Hames (2000) suggest that fertility inheritance may be larger for HGPs and

we now give some clues debating this point.

It is not clear whether to expect that fertility inheritance is more likely to occur in non HGPs than in HGPs. We will briefly review some theoretical and demographical arguments that will support the first and then the second hypothesis. It sounds reasonable and it has been checked with simulations that high fertility inheritance may lead to big sibship size (Sibert et al., 2002). However such big families are not likely to occur in stationary populations. Populations in expansion are therefore good candidates for testing fertility inheritance. The Quebecian population studied by Austerlitz and Heyer (1998), for instance, has an expansion rate of 1.4 per generation during the settlement process. Anthropologists studies have shown that there has been a first demographic transition at the Pleistocene-Holocene transition (Handwerker, 1983; Roth, 1983). The HGPs had a moderate growth rate of 1.0025 per generations at the end of the Paleolithic (Dumond, 1975). This rate switched to the higher value of 1.025 during the Neolithic period (Handwerker, 1983). The main reason usually invoked is the reduction of the birth-spacing periods once the populations started to be sedentary. The high birth-spacing period was maintained mainly by a longer lactation period (Campbell and Wood, 1998). Other reasons such as postpartum sexual abstinence and childhood infanticide (which does not reduce the birthspacing time *per se* but have an obvious impact on growth rate) have also been invoked (Dumond, 1975). Among agriculturalists, having many children became an economic advantage and the rate of population growth started to grow significantly.

Since we are looking for populations that are good candidates concerning fertility inheritance, the demographic measure that is probably more important than growth rate is the variance of the numbers of offspring. Simulations have shown that fertility inheritance increases this variance (Sibert, 2002). In the other way round, an uneven distribution of the number of offsprings will enhance the genetic effets of fertility inheritance (Austerlitz and Heyer, 1998). In contrast to the growth rates, it is not clear which kind of populations have the maximum value concerning the variance of the number of offspring. Austerlitz and Heyer (1998) showed that, in the Quebecian population they considered, this variance was much higher than predicted by a Poisson distribution with the fitted mean. In contrast, Neel and Weiss(1975) claimed than the contribution of Yanomama

(Indian hunter-gatherers from Amazonia) women is close to the Poisson distribution. However, studying the same population Chagnon (1979) found really uneven distributions whose variances are at least twice as big as the means. In brief, it is not clear whether the variance of the number of offspring is different in HGPs and a review of known census data would be welcomed.

We finally turn to the arguments that support the hypothesis that fertility inheritance is stronger in HGPs. Some anthropologists claimed that these societies were quite egalitarian without being under the control of a few leaders (Runciman, 2005). Even if this claim is highly controversial (Goody, 2005), we will adopt it for raising our next argument. In HGPs, individuals from the same sibling cooperate and they build a kin network that may be helpful in case of need. This argument has been raised for explaining the correlation between sibship size and fertility in the Ache living in Paraguay (Hill and Hurtado, 1996) and the !Kung from Botswana (Draper and Hames, 2000). It has even been noticed that the fertility of the lastborn individual in the !Kung is the highest. This finding is explained by the fact that the lastborn individual has experienced during his all childhood all the benefit of a big sibship size and he wants to reproduce it (Draper and Hames, 2000). On the contrary, in some pastoralists populations from Africa, namely the Gabbra and the Kipsigies, the fact that parents has to give brideprice to their sons may weaken family with big sibship size: The more children they have, the lower the brideprice will be (Mace, 1996; Borgerhoff Mulder, 1998).

One of the strongest exemple of fertility inheritance in HGPs, namely the Yanomana, has been given by Chagnon (1979). The major preoccupation of Yanomana men is finding women, and all the fights seem to be caused by women. Since the marriage are arranged between lineages, powerful men are able to find wifes for their numerous sons because they have many female relatives to exhange. The reproductive success of an individual, and more precisely its ability to find a wife is determided by the size of its patrilineal lineage. This situation is viable in a society were ressources are abundant and where a large household would not become a burden. Our data suggests that fertility has been inherited amongst females which is not stated in Chagnon's paper (1979). However his theoretical arguments borrowed from evolutionary biology for justifying that the society is patrilineal are dubious. He states that males are not limited biologically concern-

ing their numer of offsprings. Patrilineal societies are therefore more likely to become large and will prevail over matrilineal lineages. However recent genetic data (Helgason et al., 2002) tend to reject the fact, usually assumed, that the variances of the number of offsprings is bigger in males than in females. The tree balance of Y chromosome genealogies should be investigated in order to check whether fertility inheritance is stronger amongst males than amongst females.

The last point that support the claim that fertility inheritance is stronger in HGPs is that the level of heterozygosity is weaker in these populations. Indeed we have shown that fertility inheritance decreases heterozygosity in a drastic way. However the level of heterozygosity which is induced by strong fertility inheritance is even smaller that the one observed in HGPs (see Table 4.2). This suggests that fertility correlation has been high during a limited number of generations only. Since it leaves an high signal on genetic diversity, the balance signal was not lost, while heterozygosity could have recovered.

In summary, we have devised a test based on the balance of reconstructed gene genealogies that detect fertility inheritance. The analysis of mtDNA gene genealogies suggest that fertility inheritance between females is stronger in HGPs than in non HGPs. Possibles explanations involve that a strong and large kinship nexus may confer advantages, in particuliar for finding mates, without being an handicap as it may be in a society with scarce ressources. The marriages rules as marriages between cross-cousin observed in Yanomana (Chagnon, 1979) may enhance the strength of fertility inheritance. These populations are known to have a small effective population size caused by low densities, fragmented habitat (Ray et al., 2003) or a recent bottleneck (Excoffier and Schneider, 1999). Fertility inheritance is an other cause explaining these low effective sizes (Nei and Murata, 1966). In the future, it may be interesting to devise a new test based on genetic data that do not require any tree reconstruction. Detecting and quantifying fertility inheritance is of primary importance since it strongly affects genetic drift. An unsuspected high rate of fertility correlation between females during the Paleolothic period would reduce the time estimate of origin of Homo Sapiens that has been computed from mtDNA diversity.

Chapter 5

Minimal clade size and external branch length under the neutral coalescent

Abstract

Given a sample of genes taken from a large population, we consider the neutral coalescent genealogy, and we study the theoretical and empirical distributions of the size of the smallest clade which contains a fixed gene. We show that the theoretical distribution is strongly related to a Yule distribution of parameter two, and that the empirical count statistics are asymptotically Gaussian as the number of genes grows to infinity. Then we consider external branches of the coalescent tree, and we describe their lengths. Using the infinitely-many sites model of mutation, we also describe the conditional distribution of the external branch lengths given the number of pairwise differences between a reference DNA sequence and the sequence of one closest relative in the sample.[1]

[1]Article published: Blum M.G.B. and O. François 2005. Minimal clade size and external branch length under the neutral coalescent. *Adv. in Appl. Probab.* **37** (3) 647-662

5.1 Introduction

Coalescent theory associated with molecular information has proven an invaluable tool for assessing the degree of relatedness between individuals. To this date, these tools have been successively applied by human geneticists to infer very deep relationships, of the order of hundreds of generations (Nordborg, 2001). For example, Donnelly et al. (1996) estimated the time since the most recent common ancestor of modern humans from DNA sequences of the ZFY intron.

In this article, we are concerned with the estimation of intermediate ancestry, and more specifically the relatedness of a given gene (or individual) to a sample of $n-1$ weakly related genes. The precise details of population history or geographic structure will not enter the analysis, and the evolution of the population will be assumed to be selectively neutral. In addition, non recombining genes will be considered. In other words, genetic drift will be the only factor responsible for the allelic variation within the population. Under these assumptions, the genealogy of n genes is well approximated by the coalescent model when the population size N grows to infinity (Kingman, 1982b). The approximation arises as a diffusion limit when time is measured in units of N generations. See Tavaré (2004) or Durrett (2003) for recent reviews on the subject.

The article is structured in two parts. In the first part, we give results about the numbers of relatives of the reference gene, *i.e.* the smallest number of genes (minus one) that share an ancestor with the reference gene. This subset of genes will be called a *minimal clade* of the coalescent tree. A minimal clade contains the reference gene plus the subset of its closest relatives. We shall show that the number of closest relatives follows a Yule (or Zipf) distribution of parameter 2. Given the genealogy, we shall also study the limit probability distributions of count statistics as n grows to infinity. The studied statistics correspond to the number of genes with k relatives, $k \geq 2$. We shall prove their asymptotic normality, and hence provide new statistical tests of neutral evolution based on the shape of trees (McKenzie and Steel, 2000).

In the second part, we study the length of an external branch of the coalescent tree (Fu and Li, 1993a). This length correspond to the time since the coalescence

of an arbitrary lineage with the other lineages, and consists of a natural measure of the amount of relatedness of a gene with the rest of the sample. Using the *infinitely-many sites model* of the DNA molecule, we also describe the conditional distribution of the coalescence time given the number of substitutions between a reference DNA sequence and one of its closest relative. This distribution corresponds to an explicit mixture of gamma distributions, and extends results of Tajima (1983). An application to the human Y chromosome is presented at the end of the article.

5.2 Background and Notations

Consider a sample of n genes. In the coalescent, one wishes to record information about the number of ancestors at various times and also information about which genes share common ancestors. The *ancestral process* $A_n(t)$ records the number of distinct ancestors of the sample at a time t in the past. It can be described as a continous-time Markov chain on $[n] = \{1, \ldots, n\}$ such that

$$A_n(0) = n,$$

1 is an absorbing state, and the rate of transition from k to $k-1$ is equal to

$$\lambda_k = \frac{k(k-1)}{2}, \quad k = n, \ldots, 2.$$

This means that the times (T_k), $k = n, \ldots, 2$, separating coalescence events are independent and exponentially distributed with mean $2/k(k-1)$. In the sequel, we shall also denote

$$S_k = T_n + \cdots + T_k.$$

To understand the topology of the tree, one possibility is labelling genes in the sample from the set $[n]$ and defining a random equivalence relation. In this relation, the genes i and j are in the same class at time t iff they share a common ancestor at this time. Denote by $C(t)$ the random partition obtained from the former equivalence relation at time t. According to Kingman (1982a), the process

$C(t)$ is a continuous-time Markov chain on the set of all partitions of $[n]$ (denoted $\mathcal{E}_n$) for which

$$C(0) \equiv \{\{1\}, \ldots, \{n\}\}.$$

The transition rates of this Markov chain can be described as follows

$$\text{for all } \alpha, \beta \in \mathcal{E}_n, \quad q_{\alpha\beta} = \begin{cases} 1 & \text{if } \alpha \prec \beta \\ 0 & \text{otherwise,} \end{cases}$$

where the symbol $\alpha \prec \beta$ means that α is a nested partition of β which may be obtained by merging two classes in α. The observation that 1 is an absorbing state of $A_n(t)$ means that $C(t)$ converges almost surely to $\{[n]\}$, the final state in which a single class remains. The embedded discrete-time Markov chain $\{\mathcal{C}_k\}$, $k = n, \ldots, 1$, moves through the sequence

$$\mathcal{C}_n \equiv C(0) \prec \mathcal{C}_{n-1} \prec \mathcal{C}_{n-2} \prec \ldots \prec \mathcal{C}_1 \equiv \{[n]\}$$

and has transition probabilities

$$\mathrm{P}(\mathcal{C}_{k-1} = \alpha_{k-1} \mid \mathcal{C}_k = \alpha_k) = \frac{2}{k(k-1)}, \quad k = n, \ldots, 2.$$

Transitions happen if $\alpha_k \prec \alpha_{k-1}$ and α_k has exactly k classes, otherwise the transition probability is zero.

In the coalescent tree, a clade is an equivalence class $\alpha_\star$ for some $\mathcal{C}_i$ ($i = n-1, \ldots, 1$) such that the time since the most recent common ancestor of the genes in $\alpha_\star$ is S_{i+1} exactly. In these notations, i corresponds to the number of ancestors present in the sample at the instant of coalescence.

5.3 Main results

In this article, we consider a specific realization of the random genealogy $C(t)$. This realization can hence be represented as a rooted tree with the genes at the tips and the most recent common ancestor of the sample at the root. Consider an arbitrary gene in the sample, and give to this gene the label 1. Now, we define

the coalescence time of the lineage 1 with the rest of the sample as follows

$$\tau_n = \sup\{t \geq 0\,,\ \{1\} \text{ element of } C(t)\}.$$

The random variable τ_n corresponds to the length of a so-called external branch of the genealogy (Fu and Li, 1993a).

By definition, the smallest clade which contains the gene 1 consists of the equivalence class α_1 in the partition $C(\tau_n)$ so that $\{1\} \subset \alpha_1$. This means that α_1 contains the reference lineage 1 at the instant of coalescence with the rest of the sample. In the sequel, α_1 is sometimes called the *minimal clade*. Our interest is in the size X_n of the minimal clade. Formally, X_n is then defined as

$$X_n = \#\alpha_1, \quad \alpha_1 \in C(\tau_n), \quad \{1\} \subset \alpha_1.$$

The first result in this section gives the distribution of the random variable X_n.

Theorem 7 *Let $n \geq 2$. The random variable X_n has the following probability distribution*

$$\mathrm{P}[X_n = x] = \frac{4}{(x-1)x(x+1)}, \quad x = 2, \ldots, n-1,$$

and

$$\mathrm{P}[X_n = n] = \frac{2}{n(n-1)}.$$

The limiting distribution of X_n has mode 2 and is long-tailed. This is an expected result because this behaviour is a typical feature of the number of species in a genus in traditional hierarchical phylogenetic taxonomy (see e.g. Aldous, 2001). In the Markov linear growth model or *Yule* branching process, this number is indeed distributed according to a *Yule* law

$$\mathrm{P}[X' = x] = \rho\,\Gamma(1+\rho)\frac{\Gamma(x)}{\Gamma(x+1+\rho)}, \quad x \geq 1$$

for some parameter $\rho > 0$ (Yule, 1924). As n grows to infinity, the distribution of $M_n = X_n - 1$ converges to

$$\mathrm{P}[M = m] = \frac{4}{m(m+1)(m+2)}, \quad m \geq 1,$$

and corresponds to the Yule distribution of parameter $\rho = 2$. This is noteworthy that the distribution of X_n coincide with the minimum of $M + 1$ and n, and does not depend on n except through the event $(X_n = n)$. The small discrepancy between the finite and asymptotic probability values is created by the chance that the reference lineage connects at the root of the tree. As a corollary of Theorem 7, the expected size can be computed easily. We find that

$$\mathrm{E}[X_n] = 3 - \frac{2}{n}, \quad n \geq 2,$$

and the expected value converges to 3. In addition, the variance is equal to

$$\mathrm{Var}[X_n] = 4 \sum_{i=3}^{n} \frac{1}{i} - 6 + o(1), \quad n \geq 3,$$

which is of order $\log n$, and converges to infinity rather slowly.

Given a coalescent tree with n genes, we now consider the empirical frequencies of minimal clades of fixed size

$$f_n^x = \frac{\#\{i \,:\, X_{i,n} = x\}}{n}$$

where $X_{i,n} = \#\alpha_i$ and α_i is the equivalence class which contains the gene i in $C(\tau_{i,n})$. (Here the definition of $\tau_{i,n}$ is relative to i instead of 1). The following result states that the distribution of f_n^x is approximatively Gaussian for large n.

Theorem 8 *When n goes to infinity, we have*

$$\sqrt{n}\left(f_n^x - \frac{4}{(x-1)x(x+1)}\right) \to \mathcal{N}(0, \sigma^2)$$

where $\sigma^2 = 8/45$, *if* $x = 2$, *and*

$$\sigma^2 = \frac{4(11 + 4x^4 - 27x^2)}{x(2x+1)(2x-1)(x-1)^2(x+1)^2},$$

for all $x \geq 3$.

The case $x = 2$ is a direct consequence of McKenzie and Steel's results (McKenzie and Steel, 2000) because nf_n^2 is twice the number of cherries in a coalescent or Yule tree. McKenzie and Steel's result was derived from the analogy with extended Polya urns. Theorem 8 is based on a link to recent results in theoretical computer science regarding binary search trees. Binary search trees appear as formal representations for *divide-and-conquer* algorithms (Régnier, 1989; Rösler, 1991). The proof will exploit the one-to-one correspondence between binary search trees and coalescent trees, and use the stochastic recurrence equations involved in these data structures (Aldous, 1995).

Next, a mutation process is superimposed on the coalescent tree, and we assume that DNA sequences are observed at the tips of the genealogical tree. The times at which the mutations occur are modeled as a Poisson process of constant rate $\theta/2$, for some $\theta > 0$. If a branch of the tree has length t, then the number of mutations has a Poisson distribution with mean $\theta t/2$, independently of the other branches. Among the various models that describe the mutation types, the *infinitely-many sites* model may be one of the most appropriate (Watterson, 1975). In this model, each DNA sequence consists of completely linked sites (ie, no recombination occurs). Each mutation occurs at a site of the DNA sequence that had not been mutated before, so that a new segregating site arises. The number of segregating sites corresponds to the number of substitutions of ancestral bases since the most recent common ancestor in the sample.

Under the infinitely-many sites assumption, Tajima (1983) studied a sample of size two. After observing ℓ substitutions, it follows from Bayes Theorem that the conditional distribution of the coalescence time is a Gamma distribution of shape $(1 + \ell)$ and scale $1/(1 + \theta)$, $G(1 + \ell, 1/(1 + \theta))$, where

$$G(a, \lambda)(t) = \frac{\lambda^a}{\Gamma(a)} t^{a-1} e^{-\lambda t}, \quad t \geq 0.$$

In this article, we shall describe the conditional distribution of the coalescence time τ_n given that $\Delta = \ell$ substitutions are observed. In this notations, Δ is the number of substitutions found when comparing the DNA sequence of the gene 1 and the one of a closest parent in the coalescent tree. The conditional distribution can be formulated as a mixture of Gamma distributions. For $k = 2, \ldots, n-1$, and $k \leq j \leq n$, let us denote

$$c(j,k) = \prod_{j \leq \ell \neq k \leq n} \left(\ell(\ell-1) - k(k-1)\right).$$

In addition, we set $c(n,n) = 1$. Define

$$a_k = \frac{(n-1)!(n-2)!}{\lambda_k} \sum_{j=2}^{k} \frac{c(j,k)^{-1}}{(j-2)!^2}, \quad \text{for } k = 2, \ldots, n$$

Then, we have

$$f_{\tau_n|\Delta=\ell}(t) = \sum_{k=2}^{n} a(k,\ell) G(1+\ell, \lambda_k + \theta)(t), \quad t \geq 0,$$

where

$$a(k,\ell) = \frac{a_k \mathcal{G}_{\mathrm{s}}(p_k, \ell)}{\mathrm{P}(\Delta = \ell)}, \quad \ell = 0, 1, \ldots,$$

and

$$\mathrm{P}(\Delta = \ell) = \sum_{k=2}^{n} a_k \mathcal{G}_{\mathrm{s}}(p_k, \ell), \quad \ell = 0, 1, \ldots,$$

with $\mathcal{G}_{\mathrm{s}}(p_k,)$ the shifted geometric distribution of parameter

$$p_k = \frac{1}{1 + \theta/\lambda_k}, \quad k = 2, \ldots, n.$$

5.4 The size of the minimal clade

5.4.1 Level of coalescence with the rest of the sample

The coalescence level K of the gene 1 with the rest of the sample is defined as the random variable taking values $k = n, n-1, \ldots, 2$ so that

$$K = k \quad \text{iff} \quad \tau_n = S_k.$$

More specifically, K is related to the ancestral process as follows

$$K = 1 + A_n(\tau_n).$$

We give the distribution of K below.

Proposition 1 *For $n \geq 2$, we have*

$$\mathrm{P}(K = k) = \frac{2(k-1)}{n(n-1)}, \quad \text{for all } k = 2, \ldots, n.$$

Proof. Recall that, for all $k = 2, \ldots, n$, we have

$$\mathrm{P}(\mathcal{C}_k = \alpha) = \frac{(n-k)!k!(k-1)!}{n!(n-1)!} n_1! \ldots n_k! \tag{5.1}$$

where $\alpha \in \mathcal{E}_n$ is a partition in k classes such that each class has cardinality $\#\alpha_i = n_i$ and $n_1 + \cdots + n_k = n$. In the sequel, we shall use the notation $|\alpha| = k$. For a proof of this classical result, see Tavaré (2004, p.39, proposition 2.2.2.)

For $k = 2, \ldots, n$, the probability that the coalescence occurs at a level of the genealogy lower than k is equal to

$$\mathrm{P}(K \leq k) = \mathrm{P}(\{1\} \in \mathcal{C}_k).$$

To compute the probability of this event, we can write

$$\mathrm{P}(K \leq k) = \sum_{\alpha:|\alpha|=k-1} \mathrm{P}(\mathcal{C}_k = \{1\} \cup \alpha)$$

where the sum runs over all partitions of $\{2, \ldots, n\}$ into $(k-1)$ classes. Using equation (5.1), we find that

$$\mathrm{P}(K \leq k) = \sum_{\alpha:|\alpha|=k-1} \frac{(n-k)!k!(k-1)!}{n!(n-1)!} n_1! \ldots n_{k-1}!$$

and since $n_1 + \cdots + n_{k-1} = n-1$, we have

$$\sum_{\alpha:|\alpha|=k-1} \frac{(n-k)!(k-1)!(k-2)!}{(n-1)!(n-2)!} n_1! \ldots n_{k-1}! = 1.$$

Therefore, we obtain that

$$\mathrm{P}(K \leq k) = \frac{k(k-1)}{n(n-1)}, \quad k = 2, \ldots, n,$$

which yields the desired result. ■

Note that the distribution of K could also be obtained in a less direct way using a result by Wiuf and Donnelly (1999, p.188). The mean and the variance of K can be computed from elementary algebra. We find that the expectation is equal to

$$\mathrm{E}[K] = \frac{2}{3}(n+1), \quad n \geq 2.$$

In order to compute the variance, recall that

$$\sum_{k=2}^{n} k^3 = \frac{1}{4}\left((n+1)^4 - 2(n+1)^2 + (n+1)^2\right) - 1.$$

Then we have

$$\mathrm{Var}[K] = \frac{1}{18}(n^2 - n - 2).$$

The interpretation is that the average level at which the coalescence occurs is closer to the tips of the tree than to the root. However the variance is relatively

large with respect to the mean for large sample sizes.

5.4.2 Minimal clade size

This Section deals with the size X_n of the minimal family of an arbitrary individual in the sample. It gives a proof that the random variable X_n has a power law distribution

$$\mathrm{P}[X_n = x] = \frac{4}{(x-1)x(x+1)}, \quad x = 2, \ldots, n-1,$$

and

$$\mathrm{P}[X_n = n] = \frac{2}{n(n-1)}.$$

Before giving a proof of the Theorem, we establish a useful combinatorial identity in the next lemma.

Lemma 1 *Let $n \geq 4$, and x an integer so that $n > x \geq 3$. We have*

$$\sum_{k=x-2}^{n-3} k(k-1)\cdots(k-x+3)(n-k-1)(n-k-2) = 2\frac{n(n-1)\cdots(n-x)}{(x-1)x(x+1)}.$$

Proof. First, use induction to prove that

$$\sum_{k=x-2}^{n-1} k(k-1)\cdots(k-x+3) = \frac{n(n-1)\cdots(n-x+2)}{x-1},$$

and then use a similar recursion argument to prove that

$$\sum_{k=x-2}^{n-2} k(k-1)\cdots(k-x+3)(n-k-1) = \frac{n(n-1)\cdots(n-x+1)}{x(x-1)}.$$

Applying the recursion again, the lemma follows from the above equations. ■

Proof of Theorem 7. Consider the coalescent starting with n lineages. A standard result in coalescent theory states that if we pick one lineage at random when there are $k \leq n$ lineages, then the probability it will contain m of the n starting lineages is

$$\mathrm{P}(M_k^n = m) = \frac{\begin{pmatrix} n-m-1 \\ k-2 \end{pmatrix}}{\begin{pmatrix} n-1 \\ k-1 \end{pmatrix}}, \quad 1 \leq m \leq n-k+1.$$

For a proof of this result, see e.g. Durrett (2003, chap 1, eq. (3.14)).

At the moment of the coalescence with the rest of the sample, the lineage of individual 1 coalesces with a random subset of size M^{n-1}. Conditional to $K = k$, this means that X_n has the same distribution as

$$X_n \sim 1 + M_{k-1}^{n-1}$$

where the coalescent starts with $(n-1)$ lineages in M_{k-1}^{n-1}. Then, for all $k = 3, \ldots, n$, we have

$$\mathrm{P}(X_n = 1 + m \mid K = k) = \frac{\begin{pmatrix} n-m-2 \\ k-3 \end{pmatrix}}{\begin{pmatrix} n-2 \\ k-2 \end{pmatrix}}, \quad m = 1, \ldots, n-k+1,$$

and, for $k = 2$, we have

$$\mathrm{P}(X_n = n \mid K = 2) = 1.$$

Then, we have

$$\mathrm{P}(X_n = n) = \frac{2}{n(n-1)},$$

and, for all $m = 1, \ldots, n-2$,

$$\mathrm{P}(X_n = 1+m) = \sum_{k=3}^{n-m+1} \frac{2(k-1)}{n(n-1)} \frac{\left(\begin{array}{c} n-m-2 \\ k-3 \end{array} \right)}{\left(\begin{array}{c} n-2 \\ k-2 \end{array} \right)}.$$

For $m = 1$, we obtain

$$\mathrm{P}(X_n = 2) = \sum_{k=3}^{n} \frac{2(k-1)(k-2)}{n(n-1)(n-2)},$$

which is equal to $2/3$. Similarly, for $n > x \geq 3$, we obtain

$$\mathrm{P}(X_n = x) = \sum_{k=x-2}^{n-3} \frac{2k(k-1)\cdots(k-x+3)(n-k-1)(n-k-2)}{n(n-1)(n-2)\cdots(n-x)}.$$

Using Lemma 1, we find that

$$\mathrm{P}(X_n = x) = \frac{4}{(x-1)x(x+1)}$$

for all $x = 2, \ldots, n-1$. ■

We turn now to the proof of Theorem 8. Given a coalescent tree, we compute the number of subtrees with x leaves (genes) and a lineage that connects to the root of the subtree. Dividing by n, this leads to f_n^x which is an unbiased estimate of the probability $\mathrm{P}(X_n = x)$, for all $x = 2, \ldots, n$. When n goes to infinity, we obtain a Gaussian central limit theorem

$$\sqrt{n}\left(f_n^x - \frac{4}{(x-1)x(x+1)} \right) \to \mathcal{N}(0, \sigma^2)$$

where $\sigma^2 = 8/45$ if $x = 2$, and

$$\sigma^2 = \frac{4(11 + 4x^4 - 27x^2)}{x(2x+1)(2x-1)(x-1)^2(x+1)^2}, \tag{5.2}$$

for all $x \geq 3$.

Proof of Theorem 8. Let $X_n^x = nf_n^x$ denote the number of minimal clades of size x. The case $x = 2$ is a direct consequence of McKenzie and Steel's results (2000) because X_n^2 is twice the number of cherries in a coalescent or Yule tree. For $x \geq 3$, the proof follows from the fact that the random variable X_n^x can be involved into a quicksort-like recurrence (Hwang and Neinninger, 2002)

$$X_n = X_{I_n} + X^*_{n-I_n} + t_n \tag{5.3}$$

where I_n is uniform over the set $\{1, \ldots, n-1\}$ and the *toll* function t_n is equal to

$$t_n = \delta_{n,x}(\delta_{I_n,1} + \delta_{I_n,n-1})$$

where δ denotes the Kronecker symbol. The expression of σ^2 was found by taking the variance in both size of equation (5.3). The induction can be solved using an analytic lemma (Hwang and Neininger, 2002, Lemma 1, p.1691) and a symbolic algebra computer package. For $n \geq 2x+1$, we obtained that $\mathrm{Var}[X_n^x] = \sigma^2 n$ with σ^2 given by the equation (5.2). The final result is a consequence of Hwang and Neininger's classification of toll functions (Hwang and Neininger, 2002, p. 1701). ■

Comments. Note that the proof of Theorem 8 contains an implicit proof of Theorem 7. Taking expectations in equation (5.3) leads to recursive identities that can also be solved using Lemma 1 in Hwang and Neininger (2002). After short calculations, the results obtained in Theorem 7 can be checked again.

5.5 Some comparisons

5.5.1 Two random genes

For sake of comparison, we will also describe the distribution of the level of coalescence for two arbitrarily chosen genes. The two genes can be labeled 1 and 2. Then let us denote their coalescence time as $\tilde{\tau}_n$. We set

$$\tilde{K} = 1 + A_n(\tilde{\tau}_n),$$

and wish to compare $\tilde{K}$ to K. A well-known result in coalescence theory is that the coalescence time $\tilde{\tau}_n$ of two lineages has exponential distribution $\mathcal{E}(1)$. A next section will describe the result for τ_n. As far as the coalescence of two random lineages in the sample is concerned, we have the following probability distribution. For $n \geq 2$, we have

$$\mathrm{P}(\tilde{K} = k) = \frac{n+1}{n-1}\frac{2}{k(k+1)} \quad \text{for all } k = 2, \ldots, n.$$

The argument is based on the bivariate coalescent (see e.g. Tavare, 1996, p.87) and a result of Saunders et al. (1984) that describes the joint distribution of

$$B(t) = (A_m(t), A_n(t)), \quad t \geq 0,$$

where $A_n(t)$ is the ancestral process at time t and $A_m(t)$ is the ancestral process of a subsample of size $m \leq n$. Using the results of Saunders et al. (1984), we find that

$$\mathrm{P}(A_2(t) = 1, A_n(t) = k-1) = \mathrm{P}(A_n(t) = k-1)\ \frac{2(n-k+1)}{k(n-1)}, \quad n \geq k \geq 2.$$

For $2 \leq k \leq n-1$, we have

$$\mathrm{P}(\tilde{\tau}_n \leq S_k) = \Pr(A_2(S_k) = 1) \equiv F(k) = \frac{2}{k}\frac{n-k+1}{n-1}$$

and then

$$\mathrm{P}(\tilde{K} = k) = \mathrm{P}(\tilde{\tau}_n = S_k) = F(k) - F(k+1) = \frac{n+1}{n-1}\frac{2}{k(k+1)}.$$

For $k = n$, we have

$$\mathrm{P}(\tilde{K} = n) = \mathrm{P}(\tilde{\tau}_n \leq S_n) = F(n) = \frac{2}{n(n-1)}.$$

Note that the probability that two individuals share their most recent ancestor with the whole sample was calculated by Watterson (1982) whose result agrees

with the fact that

$$\mathrm{P}(\tilde{K}=2)=\frac{1}{3}\frac{n+1}{n-1}.$$

In conclusion, the distribution of $\tilde{K}$ is very different from K. The nodes close to the root are given more important weights in the distribution of K. This can also be seen with the average level $\mathrm{E}[\tilde{K}]$ which is $O(\log n)$ in contrast with the $O(n)$ result obtained for $\mathrm{E}[K]$.

5.5.2 A random clade

A useful comparison to Theorem 7 may be given by the distribution of the number of individuals in a random clade of the coalescent tree. Choose $I=i$ from the uniform distribution on $[n-1]$ and consider the number of genes Y_n in the (unique) clade of $\mathcal{C}_I$. We have the following result.

Proposition 2 *Let $n\geq 2$ and Y_n be the number of individuals in a random clade of the coalescent. We have*

$$\mathrm{P}(Y_n=y)=\frac{n}{(n-1)}\frac{2}{y(y+1)},\quad \textit{for all } y=2,\ldots,n-1,$$

and

$$\mathrm{P}(Y_n=n)=\frac{1}{n-1}.$$

Proof. Using arguments similar to those of Theorem 7, we deduce the conditional distribution of Y_n given that $I=i$, for $i=2,\ldots,n-1$. We obtain

$$\mathrm{P}(Y_n=y\mid I=i)=(y-1)\frac{\begin{pmatrix} n-y-1 \\ i-2 \end{pmatrix}}{\begin{pmatrix} n-1 \\ i \end{pmatrix}},\quad y=2,\ldots,n-i+1.$$

Hence, for $2 \leq y \leq n-2$, we have

$$\mathrm{P}(Y_n = y) = \frac{n(y-1)}{(n-1)} \sum_{j=y-1}^{n-2} \frac{(j-1)(j-2)\cdots(j-y+2)(n-j)(n-j-1)}{n(n-1)\ldots(n-y)}$$

which yields the result. ■

This distribution can be found in a different manner using results on binary search trees introduced in theoretical computer science. Devroye (1991) described the number of occurrence of subtrees of a given size. The above proposition is strongly connected to his results.

Let us remark that the average size of a random clade grows as $\log n$

$$\mathrm{E}[Y_n] = \frac{2n}{(n-1)}(H_n - 1), \quad n \geq 3,$$

where H_n is the nth harmonic number, while $\mathrm{E}[X_n]$ remains bounded by 3. The variance of Y_n grows as n. The asymptotic distribution of Y_n is given by

$$\mathrm{P}(Y = y) = \frac{2}{y(y+1)}, \quad y \geq 2,$$

which corresponds to the Yule distribution of parameter $\rho = 1$ given that the variable is greater than 2. According to Devroye's result and the correspondence with binary search trees, the frequencies of subtrees in a coalescent tree are asymptotically Gaussian for large n. Denoting f_n^y the frequency of subtrees of size y $(2 \leq y \leq n)$, we have

$$\sqrt{n}\left(f_n^y - \frac{2}{y(y+1)}\right) \to \mathcal{N}(0, \sigma^2)$$

where the convergence holds in distribution. Modifying Devroye's result, we obtain that

$$\sigma^2 = \frac{2(y-1)(4y^2-3y-4)}{y(2y+1)(2y-1)(y+1)^2},$$

for all $y \geq 2$.

5.6 External branch lengths

5.6.1 Unconditional distribution

The main result in this Section is the description of the distribution of the coalescence time τ_n. This random variable corresponds to the length of an external branch in the terminology of Fu and Li (1993a). Before giving the distribution of τ_n, we remark that the mean and variance of this random variable follow from Proposition 1. Fu and Li (1993a) provided a different proof for these results. Using the fact that

$$\tau_n = T_n + \cdots + T_K,$$

we obtain the following results.

Proposition 3 *Let $n \geq 2$. Consider the coalescence time τ_n. We have*

$$\mathrm{E}[\tau_n] = \frac{2}{n},$$

and

$$\mathrm{Var}[\tau_n] = \frac{4}{n^2}.$$

Now, recall that for $k = 2, \ldots, n-1$, and $k \leq j \leq n$, we denoted

$$c(j,k) = \prod_{j \leq \ell \neq k \leq n} \left(\ell(\ell-1) - k(k-1)\right), \quad c(n,n) = 1.$$

We have the following result.

Theorem 9 *Let $n \geq 2$. The Laplace transform of τ_n is given by*

$$L_{\tau_n}(s) = \mathrm{E}[e^{-s\tau_n}] = \sum_{k=2}^{n} a_k \frac{\lambda_k}{s+\lambda_k}, \quad s \geq 0,$$

where

$$a_k = \frac{(n-1)!(n-2)!}{\lambda_k} \sum_{j=2}^{k} \frac{c(j,k)^{-1}}{(j-2)!^2}. \tag{5.4}$$

The probability density function can be described as a mixture of exponential distributions

$$f(t) = \sum_{k=2}^{n} a_k \, \lambda_k \exp(-\lambda_k t), \quad t \geq 0.$$

Proof. According to proposition 1, we have

$$\mathrm{E}[e^{-s\tau_n}] = \sum_{k=2}^{n} \frac{2(k-1)}{n(n-1)} \prod_{j=k}^{n} \frac{\lambda_j}{s+\lambda_j}$$

Using fraction decomposition, we have

$$\prod_{j=k}^{n} \frac{1}{s+\lambda_j} = \sum_{j=k}^{n} \frac{2^{n-k-1} c(k,j)^{-1}}{s+\lambda_j}.$$

Let

$$b(k,j) = c(k,j)^{-1} \frac{(k-1)}{n(n-1)} \prod_{\ell=k}^{n} \ell(\ell-1).$$

Reordering the sums, we find that

$$\sum_{k=2}^{n} \sum_{j=k}^{n} \frac{b(k,j)}{s+\lambda_j} = \sum_{k=2}^{n} \left(\sum_{j=2}^{k} b(j,k) \right) \frac{1}{s+\lambda_k}$$

and

$$\lambda_k a_k = \sum_{j=2}^{k} b(j,k).$$

■

5.6.2 Conditional distributions

We now study the number of substitutions Δ in the DNA sequence of an arbitrary gene compared with the sequence of one closest relative. Recall that, for two arbitrary genes, the number of pairwise differences has a shifted geometric distribution of parameter $p = 1/(1+\theta)$. The mean is θ and the variance is $\theta + \theta^2$. We obtain the following result.

Proposition 4 *Let $n \geq 2$ and assume the infinitely-many sites model of mutation in the coalescent. Then, the number of substitutions Δ between a gene and one closest relative is distributed as a mixture of shifted geometric distributions*

$$\mathrm{P}(\Delta = \ell) = \sum_{k=2}^{n} a_k \mathcal{G}_{\mathrm{s}}(p_k, \ell), \quad \ell = 0, 1, \ldots,$$

where a_k is given in Theorem 9 equation (5.4),

$$\mathcal{G}_{\mathrm{s}}(p_k, \ell) = (1 - p_k)^{\ell} p_k, \quad \ell = 0, 1, \ldots,$$

and

$$p_k = \frac{1}{1 + \theta/\lambda_k}, \quad k = 2, \ldots, n.$$

Proof. Let $\theta > 0$. Conditional on $\tau_n = t$, $t \geq 0$, we have

$$\mathrm{P}(\Delta = \ell \mid \tau_n = t) = \frac{\theta^{\ell} t^{\ell}}{\ell!} e^{-\theta t}, \quad \ell = 0, 1, \ldots,$$

and then

$$\mathrm{P}(\Delta = \ell) = \frac{(-1)^{\ell} \theta^{\ell}}{\ell!} L_{\tau_n}^{(\ell)}(\theta)$$

where $L_{\tau_n}^{(\ell)}$ is the ℓth derivative of the Laplace transform L_{τ_n}. Using Theorem 9, we can conclude the proof. ■

The moments of Δ were found by Fu and Li (1993a) following a different method. We have

$$\mathrm{E}[\Delta] = \frac{2}{n}\theta$$

and

$$\mathrm{Var}[\Delta] = \frac{2}{n}\theta + \frac{4}{n^2}\theta^2.$$

(We use the fact that $\mathrm{E}[\tau_n] = \sum_{k=2}^{n} \frac{a_k}{\lambda_k} = \frac{2}{n}$ and that $\mathrm{E}[\tau_n^2] = 2\sum_{k=2}^{n} \frac{a_k}{\lambda_k^2} = \frac{8}{n^2}$.)

The conditional distribution of the coalescence time τ_n given that ℓ substitutions are observed can be deduced from the Bayes formula as follows

$$f_{\tau_n|\Delta=\ell}(t) = \frac{\theta^\ell}{\ell! \mathrm{P}(\Delta = \ell)} t^\ell e^{-\theta t} f_{\tau_n}(t), \quad t \geq 0,$$

Using proposition 4, the conditional density can be reformulated as a mixture of Gamma distributions

$$f_{\tau_n|\Delta=\ell}(t) = \sum_{k=2}^{n} a(k,\ell) G(1+\ell, \lambda_k + \theta)(t), \quad t \geq 0,$$

where, for $k = 2, \ldots, n$

$$a(k,\ell) = \frac{a_k \mathcal{G}_{\mathrm{s}}(p_k, \ell)}{\mathrm{P}(\Delta = \ell)}, \quad \ell = 0, 1, \ldots$$

Table 5.1 reports the values of $\mathrm{P}(\Delta = \ell)$ for $n = 10, 30$, $\ell = 0, \ldots, 9$ and $\theta = 1, 10$. Figure 5.1 displays the curves of f_{τ_n} and $f_{\tau_n|\Delta=\ell}$ for $\ell = 0, 2, 5$ and $\theta = 10$. Exact computations of the conditional expectation are reported in Table 5.2 for $\theta = 1, 10$ and $\ell = 0, \ldots, 10$. In order to provide numerical values, we used the following formula

$$\mathrm{E}[\tau_n \mid \Delta = \ell] = \frac{(1+\ell)}{\theta} \frac{\mathrm{P}(\Delta = \ell + 1)}{\mathrm{P}(\Delta = \ell)}, \quad \ell = 0, 1, \ldots$$

We remark that the distribution of Δ is concentrated on small integers, with a rapid decrease as the number of substitutions increases. As well, larger sample sizes lead to more concentrated distributions. On the other hand, the conditional expectations become rather independent on the sample size as the number of substitutions increases.

Δ	$\theta = 1$ $n = 10$	$\theta = 1$ $n = 30$	$\theta = 10$ $n = 10$	$\theta = 10$ $n = 30$
0	0.852	0.942	0.426	0.673
1	0.114	0.050	0.218	0.193
2	0.022	0.005	0.120	0.067
3	0.006	0.001	0.071	0.028
4	0.001	–	0.044	0.013
5	–	–	0.029	0.007
6	–	–	0.020	0.004
7	–	–	0.014	0.002
8	–	–	0.010	0.001
9	–	–	0.007	0.001
sum	0.999	0.999	0.963	0.994

Table 5.1: *Probability distribution of the number of segregating sites* Δ*. The symbol "–" indicates values below* 0.001*. The last row gives the sum of the ten probabilities.*

Δ	$\theta = 1$ $n = 10$	$\theta = 1$ $n = 30$	$\theta = 10$ $n = 10$	$\theta = 10$ $n = 30$
0	0.13	0.05	0.05	0.02
1	0.39	0.20	0.11	0.07
2	0.84	0.59	0.17	0.12
3	1.46	1.23	0.25	0.19
4	2.12	1.98	0.32	0.27
5	2.76	2.68	0.41	0.35
6	3.36	3.31	0.49	0.44
7	3.92	3.89	0.58	0.53
8	4.45	4.44	0.67	0.63
9	4.97	4.96	0.76	0.72
10	5.48	5.48	0.86	0.82

Table 5.2: *Conditional expectations of the coalescence time given the number of segregating sites* $\Delta = 0, \ldots, 10$*.*

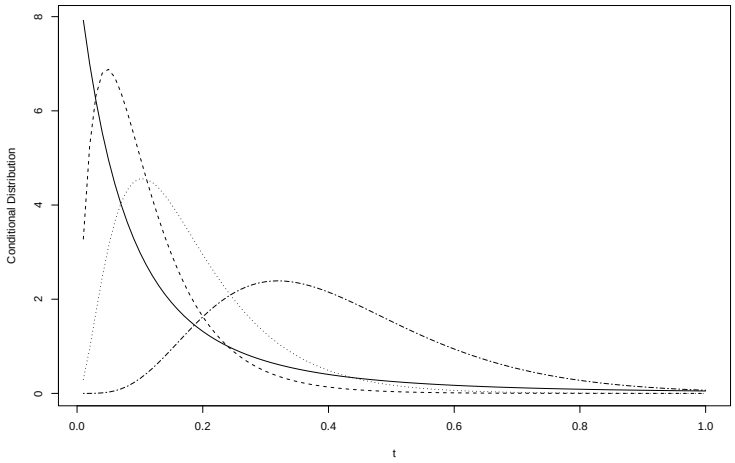

Figure 5.1: *The unconditional coalescence time distribution (solid line) and the conditional distributions given that* $\Delta = 1, 2, 5$ *segregating sites are observed. The sample size is* $n = 10$ *and the mutation rate is* $\theta = 10$.

5.6.3 Application: multilocus haplotypes

In this section, we compute posterior distributions for the time to the most recent common ancestor (MRCA) for a non recombining segment of DNA and its closest relative in a sample of $n-1$ other sequences of the same segment given that they match at ℓ out of m scored markers. The results presented here extend those of Walsh (2001) for two segments.

We consider m completely linked markers, and score their allelic states, assuming a perfect match iff no mutation occurred since the MRCA. For $\ell = 0, \ldots, m$, the conditional probability of observing ℓ matches out of m is then binomial

$$\text{for } t > 0, \quad p(\ell \mid \tau_n = t) = \frac{m!}{\ell!(m-\ell)!} e^{-\theta \ell t} \left(1 - e^{-\theta t}\right)^{m-\ell} .$$

where $e^{-\theta t}$ is the probability of a perfect match at one locus. Using Bayes formula, we obtain the conditional density of τ_n given that ℓ matches are observed

$$\text{for } t > 0, \quad p_{\tau_n}(t \mid \ell) \propto e^{-\theta \ell t} \left(1 - e^{-\theta t}\right)^{m-\ell} f_{\tau_n}(t).$$

The normalizing constant can be computed using a symbolic algebra package, or deduced from Walsh (2001). We found that

$$p(\tau_n = t \mid \ell) = \frac{\sum_{k=2}^{n} a_k \lambda_k \begin{pmatrix} m \\ \ell \end{pmatrix} e^{-(\theta \ell + \lambda_k)t} \left(1 - e^{-\theta t}\right)^{m-\ell}}{\sum_{k=2}^{n} a_k \lambda_k \begin{pmatrix} m \\ \ell \end{pmatrix} I(m, \ell, \theta, k)}$$

where

$$I(m, \ell, \theta, k) = \frac{\theta^{m-\ell}(m-\ell)!}{\prod_{i=0}^{m-\ell} \left(\lambda_k + \theta(m-i)\right)^{-1}} .$$

Given a perfect match at $\ell = m$ markers, we obtain that

$$p_{\tau_n}(t \mid m) = \frac{\sum_{k=2}^{n} a_k \lambda_k e^{-(\theta m + \lambda_k)t}}{\sum_{k=2}^{n} a_k \lambda_k (\lambda_k + \theta n)^{-1}}. \tag{5.5}$$

Walsh (2001) used a mutation rate per generation equal to $\mu = 1/500$ motivated

by estimates on the human Y chromosome, and the effective size was estimated as $N_e \approx 5,000$. These values lead to an estimate of $\theta = 20$ using that $\theta = 2N_e\mu$. Considering a perfect match at $m = 20$ markers and $n = 2$ individuals, the 95% Bayesian credible region was computed as $(6e{-}05, 0.00922)$ which corresponded to an interval of $(0.3, 46.1)$ generations for the MRCA. One conclusion was that the forensic use of the Y chromosome is rather limited (Walsh, 2001). Here, we reexamine the upper bound of the 95% Bayesian credible region given $n = 40, 60, 80, 100$ individuals. From the equation (5.5), we find that the upper bound decreases to $\lfloor tN_e + 1 \rfloor = 40, 30, 8, 7$ generations respectively. Using $m = 100$ markers, these numbers reduce to $\lfloor tN_e + 1 \rfloor = 9, 7, 5, 4$ generations.

Chapter 6

Brownian Models and Coalescent Structures

Abstract

Brownian motions on coalescent structures have a biological relevance, either as an approximation of the stepwise mutation model for microsatellites, or as a model of spatial evolution considering the locations of individuals at successive generations. We discuss estimation procedures for the dispersal parameter of a Brownian motion defined on coalescent trees. First, we consider the mean square distance unbiased estimator and compute its variance. In a second approach, we introduce a phylogenetic estimator. Given the UPGMA topology, the likelihood of the parameter is computed thanks to a new dynamical programming method. By a proper correction, an unbiased estimator is derived from the pseudomaximum of the likelihood. The last approach consists of computing the likelihood by a Markov chain Monte Carlo sampling method. In the one-dimensional Brownian motion, this method seems less reliable than pseudomaximum-likelihood.[1]

[1]Article published: Blum M.G.B, C. Damerval, S. Manel and O. François 2004. Brownian Models and Coalescent Structures. *Theor. Pop. Biol.* **65** (3) 249-261.

6.1 Introduction

In this article we discuss estimation procedures for the parameter of a Brownian motion model defined on coalescent trees. Let us consider n random variables X_1, X_2, ..., X_n, resulting at the tips of a binary rooted tree from one-dimensional Brownian random walks on the branches of this tree. We assume that the trees are randomly sampled according to Kingman's model of coalescence (Kingman, 1982a). Therefore, branch length corresponds to the time elapsed since the divergence of lineages in a neutral evolution. The random walk starts at the root of the tree, and splits into independent copies when it goes through a node. We assume that the two copies are conditionally independent given the common value at the split node. The second order structure of Brownian motions (B_t) is specified as follows

$$E[B_t^2] = \theta t, \qquad t > 0$$

for some parameter $\theta > 0$ which is the object of the estimation procedures.

Brownian motion on coalescent trees were introduced as an approximation to the ladder model of microsatellite evolution and then implemented in a computer program by Beerli (2002). This approximation replaces the discrete stepwise mutation model of Kimura and Otha (1972) with a continuous model, and assumes that the changes in microsatellite length could be approximated by a Gaussian distribution. The approximation proved useful in the context of Markov chain Monte Carlo methods, because computations could be made many times faster. Beerli (2002) reported that it appeared to work well except when genealogies have very short branches (and gave the example of those associated with very small population sizes) on which it showed a significant upward bias.

Random walks were also introduced in the context of models of *isolation by distance* in continuous populations (Wright, 1943; Malécot, 1967) where spatial dispersal is often localized in space. This approach includes a parameter σ^2 that represents the rate of dispersal, ie, the averaged squared distance between parents and offspring. Many theoretical attempts have been made in order to describe levels of genetic differentiation in terms of this parameter (e.g., Cox and Durrett, 2002). However these analyses often relied upon a discrete model, namely the

stepping stone model of Kimura (1953). Parameter estimation methods based on the stepping stone model are discussed by Rousset (2001). Here we reexamine the inference problem from another point of view. We base parameter estimation on the separation of the spatial data (ie, the locations of individuals) and the genetic data, considering the spatial data as *non-genetic* inherited characters. More specifically, Brownian motions on coalescent trees are regarded as a model for the evolution of spatial data in a large one-dimensional habitat modulo a correct rescaling of time. In fact, our approach involves the estimation of the product of the spatial dispersal rate σ^2 times the effective population size N_e

$$\theta = \sigma^2 N_e.$$

Hence, we base the estimation of θ on spatial data only. Nevertheless, estimating σ^2 yet requires genetic data because these data are usually necessary for estimating the effective population size N_e (Beaumont, 2001).

The paper is structured as follows. At the beginning of Section 2, we present the basic assumptions on which our model is based. At the end of Section 2, we describe Brownian motion on coalescent trees as the limit of discrete stepwise models on such genealogies. Two kinds of estimation methods are studied: the first based on a pairwise statistic (Section 3) and the others based on likelihoods (Section 4). Both are relevant to traditional approaches in statistical genetics. Estimation based on likelihoods is the most recent approach, and warrants optimal properties of estimators for large sample sizes. In our context, computing likelihood is a difficult issue because this function is expressed as a high dimensional integral

$$L(\theta) = \int p(D|G)p(G|\theta)dG, \tag{6.1}$$

where $p(D/G)$ is the conditional distribution of the data given the genealogy of the sample G, and $p(G/\theta)$ is the distribution of such genealogies (see Stephens, 2001). The summation over all possible genealogies cannot be performed analytically unless the sample size remains very small. Section 6.4.2 presents a fast computational method for estimating θ from a pseudomaximum-likelihood approximation. Section 6.4.4 deals with Markov chain Monte Carlo approximations.

6.2 Models

6.2.1 Coalescent trees

Kingman's coalescent genealogies (Kingman, 1982b) are large-size limits of genealogies under the the assumption that populations reproduce according to an idealized neutral *Wright-Fisher* model. Given a sample of n individuals, the ancestral process can be defined as a continuous-time Markov chain for which the jumps correspond to the times of coalescence of ancestral lineages. Let T_{n-1} be the time since the most recent common ancestor (MRCA) in the sample, T_{n-2} the time since two distinct ancestors in the sample, $T_0 = 0$. Kingman's theory states that the durations separating coalescence events $Z_k = T_{n-k+1} - T_{n-k}$, $k = 2, \ldots, n$, are independent exponentially distributed random variables of rates $\lambda_k = k(k-1)/2$. Under the assumption that time is measured in units of N_e generations, the probability distribution of genealogies G can be described as

$$p(G) = \prod_{k=2}^{n} \exp(-\lambda_k z_k). \tag{6.2}$$

An alternative way of measuring time is by rescaling as follows

$$t \equiv \theta t$$

for some $\theta > 0$. Under this transformation, the distribution of genealogies depends on the parameter θ as follows

$$p(G|\theta) = \frac{1}{\theta^{n-1}} \exp(-\sum_{k=2}^{n} \lambda_k \frac{z_k}{\theta}). \tag{6.3}$$

According to Felsenstein et al. (1999), the approximation of discrete genealogies is valid when $n^2 \ll N_e$, and was observed as being extraordinary accurate in practice.

6.2.2 Random walks

The model of Otha and Kimura (1972) is a random walk model that has been applied to the evolution of microsatellites. Microsatellites are genetic markers where a given motif of DNA is repeated several times. These data are particularly useful for population genetics studies as they are abundant and widely dispersed in eukaryotic genomes, and have high mutation rates. The number of repetitions is called the length of the microsatellite.

In the discrete ladder model of Kimura and Ohta, the population at generation ℓ consists of N_e diploid individuals. At generation $\ell + 1$, N_e offspring are created by sampling with replacement from the parental population and the parent allele can mutate with probability μ. Mutations randomly decrease or increase the length of the sequence. It is standard practice to set $\theta = 4N_e\mu$. However, we use the notation that $\theta = 2N_e\mu$. in order to obtain results homogenous with the spatial applications of Brownian motions. In the large-size limit, mutations occur according to independent Poisson processes of rate θ on each branch of the genealogy.

In a spatial model, we consider a haploid population. Each individual gives birth to a Poisson random number of offspring at rate $\lambda > 0$. Conditional to the fact that the population size is constant, this description is independent of λ, and equivalent to the Wright-Fisher haploid neutral model (Tavaré, 2001). We assume that the offspring locations are random independent variables centered around the parent location with a variance equal to σ^2. Whatever the genetic structure of the population, spatial data are therefore inherited in the same way that neutral markers could be. However, the term *neutral* is misleading as far as spatial locations are concerned. In this context, this term merely indicates the absence of density regulation. The important property is that coalescent approximation apply to this framework.

With time t measured in units of N_e generations, we set

$$\ell = \lfloor N_e t \rfloor.$$

The displacement of the offspring from an ancestor ℓ generations ago is

$$X_t = \xi_1 + ... + \xi_{\lfloor N_e t \rfloor},$$

where ξ_i corresponds to the displacement of the offspring from the parent in a single generation. In this situation, we take

$$\theta = \sigma^2 N_e,$$

and X_t has expectation $E[X_t] = 0$ and variance $Var[X_t] = \theta t$. Considering the change of variable $t \equiv \theta t$ will be useful in likelihood computations. Under this transformation, the spatial diffusion rescales to the standard Brownian process independent of θ. This change of variable is only used in section 6.4.

6.2.3 Brownian motion as an approximate model of stepwise mutation

In this section, we present informal arguments that motivate the use of Brownian models as limits of stepwise mutation models. We refer the reader to the Appendix for a more rigorous proof that Brownian models arise as the limit of sequences of compound Poisson processes which include the ladder model. Beerli (2002) reports that this kind of approximation breaks down when the effective population size is small ($\theta < 5$).

The stepwise mutation process is usually defined as follows. Let (M_t) count the number of mutations of the sequence before the time t. Mathematically, this process is defined as a homogeneous Poisson process of rate $\theta > 0$. The length variation of microsatellite markers at time t is given by

$$X_t^1 = \sum_{i=1}^{M_t} \xi_i$$

where the ξ_i are independent identically distributed discrete random variables such that

$$E[\xi_i] = 0,$$

and

$$Var[\xi_i] = \nu^2, \quad \nu > 0.$$

The basic idea that underpins the Brownian approximation is that the process (X_t^1) has the same covariance structure as the Brownian motion (B_t). A classical example of the stepwise mutation model is the symmetric random walk model

$$P(\xi_i = 1) = P(\xi_i = -1) = \frac{1}{2},$$

for which $\nu = 1$. For this model, let $s < t$, and compute

$$k(s,t) = cov(X_s^1, X_t^1) = E\left[\sum_{i=1}^{M_s} \xi_i \sum_{j=1}^{M_t} \xi_j\right].$$

Using the fact that the Poisson process has independent increments, we obtain that

$$k(s,t) = E\left[\left(\sum_{i=1}^{M_s} \xi_i\right)^2\right] + E\left[\sum_{i=1}^{M_s} \xi_i\right] E\left[\sum_{j=M_s+1}^{M_t - M_s} \xi_j\right],$$

and

$$k(s,t) = E[M_s] = \theta s.$$

For large t, M_t is equivalent to θt almost surely. According to the Central Limit Theorem, X_t^1 behaves as a Gaussian random variable $\mathcal{N}(0, \theta t)$. In addition, we have

$$cov(B_s, B_t) = \theta \min(s,t).$$

Since B_t is a Gaussian process, these equations tell that X_t^1 could be approximated by B_t.

Nevertheless, (X_t^1) is a continuous-time jump process that proceeds with discrete jumps. In contrast, (B_t) has continuous trajectories. In order to make rigorous statements, we need to rescale the processes (M_t) and (X_t^1) so that the mutations occur according to a Poisson process of rate θp. In addition, the basic jump of the rescaled process should be $\pm 1/\sqrt{p}$ instead of ± 1.

In this situation, a basic step ± 1 is the result of several steps of magnitude $\pm 1/\sqrt{p}$ which occur at rate p. When p goes to infinity, the rescaled process (X_t^p)

converges to B_t where (B_t) is $\sqrt{\theta}$ times the standard Brownian motion.

6.3 Estimation based on pairwise statistics

In this section, we investigate the properties of an estimator of θ based on pairwise statistics. Consider a dataset $D = X_1, \ldots, X_n$. The estimator is based on squared distance, as proposed by Slatkin (1995) and Goldstein et al. (1995) in the case of the stepwise mutation model. The idea behind such an estimator relies on the fact that the squared distance increases linearly with time when going forward in the genealogy.

6.3.1 Basic results about X_n

The dataset D is made of exchangeable variables, i.e., the X_i are identically distributed and the distribution of D is unchanged under arbitrary permutations of the variables. Because we assume that Brownian motions start from zero, the mean value of X_n is

$$E[X_n] = 0.$$

This can actually be shifted to any other value $E[X_n] = m$ by modifying the ancestral position from 0 to m. This may be important to do so in order to avoid negative values, in particular when microsatellite evolution is studied. The variance of X_n can be computed without difficulties

$$E[X_n^2] = \int_0^\infty E[B_t^2] f_{T_{n-1}}(t) dt = \theta E[T_{n-1}] = 2\theta(1 - 1/n)$$

and we see that an upper bound is 2θ (The averaged squared distance between clumps stay bounded away).

6.3.2 Squared distances

The distance between X_i and X_j is defined as follows

$$d_{ij} = |X_i - X_j|,$$

and we study the pairwise squared distance statistics

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_i X_i^2 - (\bar{X}_n)^2\right).$$

Moments In this paragraph, we describe the moments of the difference $X_1 - X_2$ for two randomly chosen variables X_1 and X_2 in D. Let T be the time until the most recent common ancestor of the two individuals. Given $T = t$, $X_1 - X_2$ follows a Gaussian distribution of mean 0 and variance $2\theta t$. So, we have

$$\begin{aligned} E[X_1 - X_2|T] &= E[(X_1 - X_2)^3|T] = 0 \\ E[(X_1 - X_2)^2|T] &= E[d_{12}^2|T] = 2\theta T \\ E[(X_1 - X_2)^4|T] &= E[d_{12}^4|T] = 12\theta^2 T^2. \end{aligned}$$

Because T follows an exponential distribution of parameter 1, we have

$$\begin{aligned} E[d_{12}^2] &= \int_0^\infty E[S_n^2|T=t] f_T(t)\, dt = 2\theta \int_0^\infty t f_T(t)\, dt \\ &= 2\theta E[T] \\ &= 2\theta. \end{aligned} \tag{6.4}$$

The fourth order moment of d_{12} can be computed as follows

$$\begin{aligned} E[d_{12}^4] &= \int_0^\infty E[d_{12}^4|T=t] f_T(t)\, dt = 12\theta^2 \int_0^\infty t^2 f_T(t)\, dt \\ &= 12\theta^2 E[T^2] \\ &= 24\theta^2. \end{aligned} \tag{6.5}$$

Then, the variance of the squared distance is

$$\begin{aligned} Var[d_{12}^2] &= E[d_{12}^4] - E[d_{12}^2]^2 \\ &= 24\theta^2 - 4\theta^2 \\ &= 20\theta^2. \end{aligned} \tag{6.6}$$

Pritchard and Feldman (1996) obtained similar equations in the case of the stepwise mutation model. Nevertheless, their result regarding the fourth order moment $E[d_{12}^4]$ was different and involved an additional 2θ. To conclude this paragraph, we remark that the distribution of d_{12} has a very simple expression

$$\begin{aligned} P(d_{12} > t) &= \int_{s=0}^{\infty} P(|B_{2s}| > t) e^{-s}\, ds \\ &= \exp(-t/\sqrt{\theta}), \quad t > 0, \end{aligned}$$

from which the moments could be deduced again.

Bias The result is that S_n^2 is an unbiased estimator of θ. Indeed, we have

$$S_n^2 = \frac{1}{(n-1)n} \sum_{i<j} d_{ij}^2,$$

and since the X_i's are exchangeable, we find

$$E[S_n^2] = \frac{1}{2} E[d_{12}^2] = \theta.$$

Variance In order to find the variance of S_n^2, we follow the same lines of proof as Pritchard and Feldman (1996). Their computations were based on the second and fourth moments of d_{12}, for which we obtained explicit expressions in a previous paragraph. The variance of S_n^2 is

$$Var[S_n^2] = \frac{2\theta^2(1 + 3n + 2n^2)}{3(n^2 - n)}$$

Note that S_n^2 is not a consistent estimator of θ

$$Var[S_n^2] \to \frac{4}{3}\theta^2, \quad \text{as } n \to \infty.$$

Remark Define $(X_1^p, X_2^p, .., X_n^p)$ as being a dataset obtained at the leaves of a coalescent tree from the dynamics described in Section 2 and in the appendix. Assume the convergence of the moments of $(X_1^p, X_2^p, .., X_n^p)$ to those of $(X_1, X_2, .., X_n)$.

Let $d_{ij}^{(p)}$ be the difference between X_i^p and X_j^p. Define

$$\begin{aligned} S_n^{2(p)} &= \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} d_{ij}^{(p)2} \\ &= \frac{1}{pn(n-1)} \sum_{1 \le i < j \le n} (\sqrt{p}\, d_{ij}^{(p)})^2 \end{aligned}$$

For Bernoulli random walks, the random variable $\sqrt{p}\, d_{ij}^{(p)}$ has the same distribution as the difference in repeat number between two individuals in a single step mutation model where the mutation rate is equal to $p\theta$. Applying the result of Pritchard and Feldman (1996), we have

$$Var[S_n^{2(p)}] = \frac{1}{p^2} \frac{\theta p[n(n+1)] + 2\theta^2 p^2[1 + 3n + 2n^2]}{3(n^2 - n)}$$

Thanks to the convergence of moments, the variance of S_n^2 is

$$Var[S_n^2] = \lim_{p \to \infty} Var[S_n^{2(p)}] = \frac{2\theta^2(1 + 3n + 2n^2)}{3(n^2 - n)}$$

and this establishes a direct proof of the result.

6.4 Estimation based on likelihood

6.4.1 A peeling algorithm

Given the set of data $D = x_1, \ldots, x_n$, likelihoods can be computed as the integral of $p(D/G) \times p(G/\theta)$ over all possible neutral genealogies G according to equation (6.1). Kingman's formula gives the distribution of genealogies $p(G/\theta)$ (see Section 2). In this Section, we describe a peeling algorithm that enables computing the conditional distribution $p(D/G)$ given the genealogy analytically. This procedure is based on the explicit calculation of the integrals that arise at each internal node of the tree when applying Felsenstein's likelihood method (Felsenstein, 1981).

As usual in phylogenetic likelihood algorithms, we associate trees with $(n - 1) \times 4$ arrays as follows

row i	node 1	node 2	ancestor	coalescence time

$(i = 1, \ldots, n-1)$ where n is the sample size. In addition, we assume that the coalescence times are ranked in increasing order, i.e., $t_1 < t_2 < \ldots < t_{n-1}$. The leaves are labelled from 1 to n and the other nodes (corresponding to the ancestors) are labelled from $n+1$ to $2n-1$. For example, consider the tree with 4 leaves given by

$$G = \begin{array}{ccc||c} 1 & 2 & 5 & t1 \\ 3 & 5 & 6 & t2 \\ 4 & 6 & 7 & t3 \end{array}$$

For $D = x_1, x_2, x_3, x_4$, Felsenstein's method computes $p(D|G)$ in the following way

$$\begin{aligned} p(D|G) = \int \pi(x_7) p(x_4|x_7) \quad & \times \int p(x_6|x_7) p(x_3|x_6) \\ & \times \int p(x_5|x_6) p(x_1|x_5) p(x_2|x_5)\, dx_5 dx_6 dx_7. \end{aligned} \tag{6.7}$$

where π is the distribution of $X_7 \equiv X_{MRCA}$ the value taken by the MRCA. Here, we consider a degenerate distribution where

$$X_{MRCA} = m,$$

for some m in $\mathbb{R}$. The procedure could be extended to arbitrary Gaussian distributions without difficulties. The computation of integrals of the type described above can be performed using the following technical result. Let $\alpha > 0$ and β, γ arbitrary real numbers, we have,

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2 + 2\beta x - \gamma} dx = \sqrt{\frac{\pi}{\alpha}} \exp(-\gamma + \beta^2/\alpha).$$

The distributions at each node can be characterized by four parameters A, α, β, γ, i.e.,

$$p(x \mid A, \alpha, \beta, \gamma) = A \exp(-\alpha x^2 + 2\beta x - \gamma)$$

We are now ready to describe the successive steps of the peeling algorithm.

1. The peeling algorithm is initialized at the leaves of the tree as follows. If node n_1 corresponds to a leaf, then we take

$$A = \frac{1}{\sqrt{2\pi z}} \quad \alpha = \frac{1}{2z} \quad \beta = x_{n_1}\alpha \quad \gamma = x_{n_1}^2\alpha$$

where z is the time until the most recent common ancestor with another node n_2 (z can be read in the fourth column of the tree structure). This computation corresponds for instance to the parameters that define $p(x_1|x_5)$ and $p(x_2|x_5)$ in the example genealogy G.

2. Parameters at internal nodes, e.g., corresponding to the integral

$$I = \int p(x_5|x_6)p(x_1|x_5)p(x_2|x_5)dx_5,$$

are computed recursively thanks to the following induction formula

$$\begin{aligned} A &= A_1A_2/\sqrt{2z(\alpha_1+\alpha_2)+1} \\ \alpha &= \tfrac{\alpha_1+\alpha_2}{2z(\alpha_1+\alpha_2)+1} \\ \beta &= \tfrac{\beta_1+\beta_2}{2z(\alpha_1+\alpha_2)+1} \\ \gamma &= \gamma_1+\gamma_2-\tfrac{2z(\beta_1+\beta_2)^2}{2z(\alpha_1+\alpha_2)+1} \end{aligned}$$

where $(A_1, \alpha_1, \beta_1, \gamma_1)$ and $(A_2, \alpha_2, \beta_2, \gamma_2)$ are the parameters at offspring nodes n_1 and n_2 and z is the time since the ancestor of the internal node.

3. The algorithm terminates with an integral at the root of the tree, for which we use the same formula with $z = 0$. Finally it returns

$$p(D|G) = A\exp(-\alpha m^2 + 2\beta m - \gamma)$$

where m is the mean of the sample.

A straightforward modification of this algorithm allows computing the log of distributions $\log P(D|G)$ instead of $P(D|G)$. The set of recursively computed pa-

rameters is then $\log A$, α, β and γ. In practice, we set $m = \bar{x}$ (the mean of the observed data). If a Gaussian distribution is assumed at the root of the tree, the final step of the algorithm could also use

$$z = \bar{s^2}(1 - 1/n)$$

with $\bar{s^2}$ the empirical variance of the data. In simulation experiments, we assumed a deterministic value at the root of the genealogy.

6.4.2 Pseudomaximum-likelihood Algorithm

When computing $L(\theta)$, the genealogy of the data is unknown. Nevertheless, this genealogy may be viewed as an hidden or latent random variable over which an average must be performed. A reasonable guess of what the hidden variable looks like can help computing efficient numerical approximations of the likelihood. Techniques that employ such approximations are often called pseudomaximum-likelihood methods (Seo et al., 2002).

In this section, we build a pseudomaximum-likelihood method for estimating the parameter θ based on a specific genealogy. This tree is built from a phylogenetic reconstruction method that uses squared Euclidean distances between taxa (Slatkin, 1995). Because we make the hypothesis of constant evolution along the lineages, the UPGMA (Unweighted Pair Group Method using Averages) method is a natural mean to construct the topology of the tree (Nei, 1987).

Given the UPGMA topology, the lengths of the branches are taken equal to the average intercoalescence times, $z_n = 2\theta/n(n-1)$, ..., $z_3 = 3\theta$, $z_2 = \theta$. Therefore, likelihoods are computed according to the peeling algorithm of section 6.4.1 within an $O(n)$ time. The pseudomaximum-likelihood parameter $\hat{\theta}$ is then estimated thanks to a dichotomic search method. A limitation is that the method leads to a strong downward bias for large θ. This bias is due to the fact that squared distances do not account for the large deviations of Brownian motions.

In a first stage, we investigate the way of correcting the bias of the estimator using a Monte Carlo study for different parameter settings. In these experiments, datasets are created using coalescent simulations. For a given true parameter θ,

a genealogy is sampled. This genealogy is then used to evolve the node variables according to the Brownian motion model. The data resulting at the tips of the simulated tree are then exploited in order to study the properties of the estimator. The experimental design consists of 1,000 repetitions for each parameter setting (θ, n). The parameter θ ranges from 0.1 to 50, and the sample size ranges from $n = 10$ to $n = 1000$.

Bias The main result of the simulation study is that the pseudomaximum-likelihood estimator $\hat{\theta}_n$ exhibits a constant multiplicative bias

$$E[\hat{\theta}_n] \approx b_n \theta.$$

In this relationship, the parameter b_n depends on the sample size n and is independent on θ (Table 6.1). This observation is consistent with the scaling property of Brownian motions. The values of the coefficient b_n are estimated as the correlation coefficient of a linear regression. In addition, a second regression analysis shows that the equation

$$E[\hat{\theta}] = \exp(-0.611 + 0.005n - 0.248\sqrt{n})\theta$$

fits the relationship between the average value of $\hat{\theta}$, θ and n extremely well ($R^2 \approx 0.99$). This formula provides a systematic way of correcting the estimation bias.

n	b_n	Std.Error	$t-$value	$Pr(>\|t\|)$	R^2
10	0.277	0.0049	56.6	1.34e−29	0.99
20	0.202	0.0034	58.1	6.58e−30	0.99
30	0.165	0.0018	87.2	1.22e−34	0.99
40	0.138	0.0019	70.8	3.24e−32	0.99
50	0.119	0.0015	74.8	7.62e−33	0.99
100	0.074	0.0007	101.4	2.09e−36	0.99
150	0.057	0.0010	53.1	7.38e−29	0.99
200	0.046	0.0005	82.6	5.23e−34	0.99
250	0.040	0.0003	115.4	6.38e−38	0.99
300	0.035	0.0004	78.8	1.84e−33	0.99
350	0.031	0.0004	70.7	3.44e−32	0.99
400	0.033	0.0012	26.1	1.07e−20	0.96
450	0.028	0.0005	54.1	2.86e−28	0.99
500	0.028	0.0010	27.8	1.91e−21	0.96
1000	0.014	0.0001	95.5	1.20e−34	0.99

Table 6.1: Linear regression results for the multiplicative bias of $\hat{\theta}$, $E[\hat{\theta}_n] = b_n\theta + a_n$. The parameter θ ranges from 0.1 to 50. The linear coefficients are computed for different sample sizes. The intercepts a_n are not significant.

Variance In a second stage, we investigate the quality of the pseudomaximum-likelihood estimator after bias correction. The corrected estimator is computed as

$$\tilde{\theta}_n = \frac{\hat{\theta}_n}{b_n} = e^{0.611-0.005n+0.248\sqrt{n}}\hat{\theta}_n$$

For sample sizes below $n = 300$, a quadratic relationship between θ and the variance of $\tilde{\theta}$ can be identified thanks to a regression method

$$Var(\tilde{\theta}_n) = d_n\theta^2.$$

Table 6.2 reports the values of the quadratic coefficient d_n and the significance levels of both coefficients c_n and d_n in the regression model $Var(\tilde{\theta}_n) = c_n\theta + d_n\theta^2$. In Table 6.3, we report the average values of $\tilde{\theta}$ and the standard deviations of this estimator. The last column in Table 6.3 gives the standard deviations of the unbiaised pairwise estimator s_n^2. For small values of θ ($\theta \leq 5$), and samples of intermediate size (about $100 - 500$) individuals the pseudomaximum-likelihood estimator is significantly better than the pairwise statistic. This observation remains true for larger θ, but the relative benefit is slighlty lower.

n	c_n	StdError	$t$value	$Pr(>\|t\|)$	d_n	StdError	$t$value	$Pr(>\|t\|)$
20	−0.276	0.223	−1.237	0.233	0.230	0.038	5.960	1.55e−05
30	0.179	0.107	1.660	0.115	0.075	0.0186	4.041	8.4e−04
40	0.068	0.222	0.308	0.762	0.058	0.038	1.536	0.143
100	−0.038	0.042	−0.911	0.375	0.041	0.007	5.604	3.16e−05
150	0.0008	0.046	−0.018	0.985	0.025	0.007	3.201	0.005
200	0.013	0.024	0.527	0.605	0.014	0.004	3.387	0.003

Table 6.2: Regression results for the variance of the pseudomaximum-likelihood estimator $Var[\tilde{\theta}] = c_n\theta + d_n\theta^2$. The linear coefficients c_n are nonsignificant.

θ	n	mean($\tilde{\theta}$)	sd($\tilde{\theta}$)	sd(s_n^2)	θ	n	mean($\tilde{\theta}$)	sd($\tilde{\theta}$)	sd(s_n^2)
0.1	20	0.104	0.080	0.122	0.5	20	0.522	0.401	0.614
	50	0.103	0.080	0.118		50	0.515	0.400	0.591
	100	0.095	0.067	0.116		100	0.476	0.337	0.584
	200	0.100	0.046	0.116		200	0.500	0.231	0.580
	300	0.091	0.047	0.115		300	0.459	0.237	0.579
	500	0.086	0.050	0.115		500	0.444	0.260	0.578
	1000	0.105	0.099	0.115		1000	0.467	0.273	0.578
1	20	1.086	0.850	1.229	5	20	4.738	4.251	6.145
	50	0.945	0.658	1.183		50	4.799	3.526	5.919
	100	1.078	0.680	1.169		100	4.832	3.443	5.846
	200	0.989	0.758	1.161		200	5.824	3.595	5.809
	300	1.038	0.778	1.159		300	4.592	2.548	5.797
	500	0.737	0.480	1.157		500	6.500	3.337	5.787
	1000	1.272	0.684	1.156		1000	4.887	2.794	5.780
10	20	11.154	7.502	12.295	50	20	48.541	38.142	61.451
	50	10.446	9.079	11.839		50	49.162	36.753	59.195
	100	10.980	8.883	11.692		100	50.951	34.421	58.460
	200	10.950	8.731	11.619		200	49.428	35.876	58.096
	300	9.020	4.851	11.595		300	51.542	26.653	57.976
	500	11.589	4.888	11.575		500	52.012	24.651	57.879
	1000	10.01	5.251	11.561		1000	48.013	29.651	57.807

Table 6.3: Bias and variance of pseudomaximum-likelihood estimator after correction. Parameter settings vary from $\theta = 0.1$ to $\theta = 50$, and sample sizes vary from $n = 20$ to $n = 1000$. The last column reports the standard deviations of the unbiaised pairwise estimator.

6.4.3 Lower bound of the variance of estimators of θ

Consider the random genealogy associated with the sample of data $X_1, \ldots, X_n$. There are exactly k branch segments in the tree during the time that separates the k^{th} and the $(k-1)^{th}$ internal nodes.

As Fu and Li (1993b), we measure time in generations and not in unit of N generations (in this subsection only). Let us denote $d_{k,i}$ $(i = 1, \ldots, k)$, the algebraic distance covered by the Brownian motions during z_k generations along the ith branch. The distribution of $d_{k,i}$ is Gaussian with mean 0 and variance $\sigma^2 z_k$, where we use the notation $\theta = \sigma^2 N$.

To establish a lower bound of the variance of estimators of θ, we follow the same lines of proofs as Fu and Li (1993b). This approach consists of assuming that all evolutionary events are observable. More precisely, we consider that the true topology of the genealogy is known as well as the number of generations between coalescent events $\{z_k, k = 2 \ldots n\}$, and the algebraic distances covered by Brownian motions along each branch $\{d_{k,i}, k = 2 \ldots n, i = 1 \ldots k\}$. Such a statistical model is called the complete case by Klein et al. (1999) in the context of the infinitely-many-sites model. This model contains more informations than the incomplete model where the observations are the resulting values of the Brownian motions at the tips of the tree. Because the distribution of the topology is independent on N and σ, we can restrict ourselves to the subset of data

$$S = \{z_k, d_{k,i}; k = 2 \ldots n; i = 1 \ldots k\}.$$

Given S, the likelihood can be computed as follows

$$\begin{aligned} L(\sigma^2, N; S) &= \prod_{k=2}^{n} \left(\prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2 z_k}} e^{-\frac{d_{k,i}^2}{2\sigma^2 z_k}} \right) \frac{k(k-1)}{2N} e^{-\frac{k(k-1)z_k}{2N}} \\ &= \prod_{k=2}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2 z_k}} \right)^k e^{-\frac{d_k}{2\sigma^2 z_k}} \frac{k(k-1)}{2N} e^{-\frac{k(k-1)z_k}{2N}} \end{aligned} \tag{6.8}$$

where

$$d_k = \sum_{i=1}^{k} d_{k,i}^2.$$

The loglikelihood is

$$\begin{aligned}\log L &= C - \frac{(n-1)(n+2)}{4}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{k=2}^{n}\frac{d_k}{z_k} \\ &\quad -(n-1)\log N - \sum_{k=2}^{n}\frac{k(k-1)z_k}{2N} \qquad (6.9)\end{aligned}$$

where C is independent of N and σ. The Fisher's information matrix can be deduced from equation (6.9) by computing the second order derivatives and taking the opposite of their expected values

$$I(N,\sigma^2) = \begin{pmatrix} \frac{n-1}{N^2} & 0 \\ 0 & \frac{(n-1)(n+2)}{4\sigma^4} \end{pmatrix}$$

Theorem 10 *In the complete case and in the incomplete case, the variance of all unbiased estimator $\check{\theta_n}$ of θ is bounded below by*

$$Var[\check{\theta_n}] \geq \frac{(n+6)}{(n-1)(n+2)}\theta^2. \qquad (6.10)$$

This bound is asymptotically proportional to $1/n$ which is the typical variance of an estimator build according to independent observations. It can be compared to a similar bound found by Fu and Li (1993b, eq. (23)) in the infinite many sites model asymptotically equal to $1/\log n$. The difference between the two results is a consequence of the statistical properties of the Poisson process. In the infinitely-many-sites model, the quantity of information depends linearly on the time elapsed since the root of the tree. Thus, the quantity of information in the complete case is proportional to the length of the tree which is asymptotically equal to $1/\log n$. In the Brownian model, the quantity of information is constant on any branch segment because of the rescaling property of Brownian motions. Thus, the quantity of information brought by the Brownian motions along the $O(n^2)$ branch segments is proportional to n^2. The information contained in the intercoalescence times is proportional to n and gives the major contribution to the Cramer-Rao estimate.

6.4.4 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods are computationally intensive statistical methods that have proven successful in estimating parameters of population genetics models. For instance, these methods have recently been used in estimating mutation rates (Kuhner et al., 1995), gene flow parameters (Beerli and Felsenstein, 2001), or the distribution of the time since the most recent common ancestor of a population (Tavaré, 2001; Griffiths and Tavaré, 1994a).

Method description Our approach is similar to the one described in (164). In order to compute the likelihood of θ, an importance sampling method is applied. We use the Metropolis-Hastings (MH) algorithm for drawing samples of genealogies $G^1, G^2, \ldots$ from the importance distribution

$$q(G) = p(G|D) \propto p(D|G)p(G|\theta_0)$$

where θ_0 is an initial value. Our approach to importance sampling is based on the conditional coalescent. At this stage, we use a *Gibbs sampler* implementation (Robert and Casella, 1999). The strategy consists of removing a single internal node of the tree G at each proposal step, and then simulating the conditional coalescence time of this node given that the other times remain unchanged. The removed node is therefore randomly reintroduced into the tree with its new coalescence time.

Given M trees, we compute relative likelihoods as follows

$$\frac{L(\theta)}{L(\theta_0)} \approx \frac{1}{M} \sum_i \frac{p(G^i|\theta)}{p(G^i|\theta_0)}.$$

In order to suppress the dependence on θ_0, this initial setting is kept only during a preliminary sampling. Then a maximum likelihood value θ_1 based on this sample is found. A second Markov chain starts from θ_1, and a new maximum likelihood value θ_2 is found, etc. As suggested by Kuhner et al. (1995), we run 10 short chains and two longer ones at the end of the run.

Results Datasets were created according to the same procedure as the one used in Section 5.2. Table 6.4 reports simulation results regarding the convergence of the MCMC estimator $\bar{\theta}$ for sample sizes $n = 20, 50, 100, 200$ and parameters $\theta = 1, 2.5, 5$. Biases and standard deviations are reported. In these experiments, the starting value was set to $\theta_0 = 3$. For small sample sizes ($n \leq 50$), the bias appears to be low. This can be explained as the set of genealogies is correctly explored, and the Monte Carlo Markov chain reached stationarity. For $n \geq 100$, the algorithm gets stuck in local optima more frequently. The standard deviations decrease as the sample sizes increase from $n = 20$ to $n = 50$. For large sample sizes, small variances may indicate that the algorithm has difficulties of escaping from the initial settings.

In a second series of experiments, the algorithm was run several times on two different simulated datasets ($n = 100, \theta = 50$). In the first dataset, two different subpopulations are emerging whereas there are no distinct clusters in the second one (Figure 6.1). In order to reconstruct the deep branches of the tree, more information is present in the first dataset than in the second dataset. The algorithm behaves differently in the two cases. While the estimation of θ is quite good for the first dataset, it is inaccurate for the second dataset (Table 6.5). This indicates that branches close to the root may have a strong influence on the variance of the estimator.

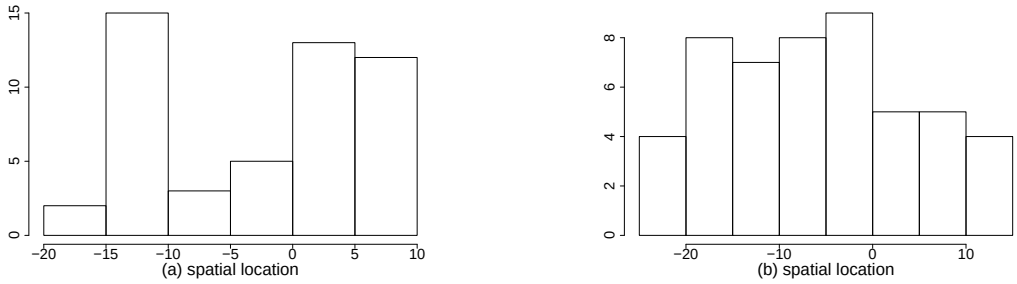

Figure 6.1: Histograms of 2 samples simulated according to the Brownian model on coalescent trees. The first data set (a) exhibits two distinct clusters whereas the spatial distribution of the second one (b) is more uniform. There are 50 resulting individuals and $\theta = 50$ in both simulations.

θ	n	mean(θ)	s.d.(θ)	5th percentile	95th percentile
1	20	.99	.29	.60	1.63
	50	1.07	.074	.57	1.63
	100	1.44	.34	.78	2.15
	200	2.95	1.06	1.31	5.04
2.5	20	2.7	1.4	1.23	6.04
	50	2.36	.45	1.52	3.29
	100	3.16	.97	1.73	5.62
	200	3.03	.66	1.89	4.29
5	20	4.2	1.16	1.70	5.88
	50	3.89	1.93	1.55	7.90
	100	3.38	.97	1.97	5.36
	200	3.25	.61	2.35	4.47

Table 6.4: Properties of the MCMC estimator $\bar{\theta}$. For each value of θ and n, mean and standard deviations are evaluated from 50 simulated datasets.

	data set 1	data set 2
mean	59.7	93.3
std.dev.	22.7	30.9
median	55	87
5th percentile	33.4	49.35
95th percentile	92.6	149.5

Table 6.5: Properties of the MCMC estimator $\bar{\theta}$, 50 estimations have been run on each data set. In data set 1, two clusters are emerging whereas no distinct clusters are seen in data set 2. The estimator is more accurate on data set 1. The true value is $\theta = 50$, the initial value θ_0 has been set up to $\theta_0 = 70$.

6.5 Spatial dispersal: Application to a biological dataset

Estimating the amount of spatial dispersion of a species, σ^2, is crucial for many ecological studies. Two very different approaches to estimating this parameter can be used: (1) direct methods using direct observations of moving individuals (e.g., via mark and recapture), and (2) indirect methods using genetic data from samples of individuals. Direct methods can help to determine the spatial pattern of dispersion during the study, and can deliver information about very recent history. However, there are also obvious shortcommings: The movements of individuals may be artefacts of the study, the accuracy of the parameter estimates may be small, and small dispersal rates may be undetectable.

Several indirect methods were devoted to the estimation of σ^2 based on molecular data. Slatkin (1993) showed that for allele frequency data, under a variety of dispersal models, there is an approximately linear log-log relationship between the product of the effective size and migration rates $N_e m_{ij}$ between a pair of populations i and j and their geographic distance, D_{ij}, namely:

$$\log N_e m_{ij} = a + b \log D_{ij}$$

A statistically significant negative regression coefficient, b, indicates that migration between populations becomes lower as their geographic separation increases, due to isolation by distance. This indicates that the regression coefficient contained information about the parameter σ^2. Rousset (1997) introduced a method based on the computation of F-statistics which exploit Slatkin's idea. He obtained an approximately linear relationship between F_{st} and the log of the geographic distance. This gave a practical method of estimating $d\sigma^2$, where d is the population density per surface unit. Rousset (2001) provides a recent survey of other methods available for estimating σ^2, and discusses the limitations of each method. For instance, the above method has the drawback of assuming well recognizable demes of several individuals, and estimators based on F_{st} may have high variance.

In Brownian models, the dispersal parameter is defined as being the standard deviation in a model where the locations of offspring followed from a probabilistic

distribution centered around the parent. Felsenstein (1975) reported the existence of clumps in similar continuous models of isolation by distance, and some regulation of the density of individuals might be added in view of realistic applications (Barton et al., 2002). Brownian motions arise naturally in continuous limits from the Kimura's stepwise migration models (Barton et al., 2002; Nagylaki, 2002). In these works, the authors usually study the conditional coalescence time given the spatial locations of two or more individuals. The typical conditional distribution has infinite mean, and hence is very different from the unconditional coalescence times used in the present article. However, although Brownian models might represent a rough approximation of the biological reality, it is useful because a number of theorical insights are available.

As an example, we illustrated our approach with a sample of spatial locations of female brown bears in Scandinavia. The Scandinavian brown bear population has a strong phylopatry of females. Hence, the spatial locations of females can be thought as being maternally inherited like a haploid character. The Scandinavian brown bear population is subdivided in two distinct populations located at the South and North of an area covering both Sweden and Norway (Taberlet et al., 1994). These two subpopulations are isolated by the distance and regulated by male migration (Waits et al., 2000). We analyzed a sample of 64 female bears locations in the South area. These data were recorded as the latitude and the longitude of individuals at the instant of capture and then converted in kilometers (km). For this dataset, the Mantel test of isolation by distance was nonsignificant, and GENEPOP gives an estimate of $\sigma = 2$ km (Rousset and Raymond, 1995), which is not in agreement with the knowledge of this population (Waits et al., 2000). The estimation of the effective size N_e is a critical step if one is interested in estimating σ^2 from θ. In this step, molecular data play an important role. We used as an estimate $N_e \approx 50$ (see Waits et al., 2000), confirmed by a multilocus microsatellite study that gave an expected homozygosity about 0.5 in this geographical area (the mutation rate can be taken as $\mu \approx$ 10e-2 (Paetkau et al., 1995)). We obtain $\theta_x = \sigma^2 N_e \approx 4350$ and $\theta_y \approx 3034$ which means that $\sigma \approx 9 - 10$ km which is more consistent with field observations (Eva Bellemain, private comm.).

6.6 Discussion

In this article, Brownian motions were considered as models of evolution of genetic data (microsatellite) as well as models for the inheritance of non-genetic features (spatial locations). We proposed three estimators for the parameter of such models. The first estimator was based on mean pairwise square distance. The second estimator was a phylogenetic estimator relying on the UPGMA topology and mean coalescence times. The third estimator was based on approximate maximum likelihood using MCMC methods.

We found the exact variance of the mean pairwise estimator. Regarding the phylogenetic estimator, we found a systematic way of correcting the biases. After the correction, the quality of the estimation improves significantly. In addition, this approach has the merit of being very fast (few milliseconds runtimes for sizes of several hundred data). The MCMC method does not lead to improved estimation. In addition, the practical implementation of the MCMC method raises a number of questions that are specific to this family of algorithms. For instance, Wilson and Balding (1998) also reported biases for MCMC estimators in the context of microsatelitte data when a single locus is used and evolution is modeled as a discrete random walk. This is in agreement with our results which show that the most likely genealogies can hardly be sampled using the information contained in a one-dimensional random walk.

Regarding the convergence issue of MCMC, some difficult problems remain to be solved, where the relevance of the transition kernel is of primary importance. Even if our transition kernel had all the theoretical properties required, many other kernels should be tested and the choice for an optimal one is an open question. MCMC is time consuming. Theoretically it is asymptotically unbiaised, but our simulations shows the difficulty to tune the several internal parameters of the MCMC algorithm.

Some methods based on coalescent theory enable the estimation of the effective size N_e in recently isolated genetically diverging populations (O'Ryan et al., 1998). These approaches require the knowledge of an additional event: the time since the (usually two) populations have been isolated from each other. In the same spirit

as (O'Ryan et al., 1998), our approach also provides an estimator of N_e given an estimator of σ^2 based on spatial data. For instance, indirect estimation of σ^2 using DNA fingerprinting (Bossart and Prowell, 1998) aims to exploit the recent shared genetic history between parents and offspring. Rousset's approach could be utilized as well, although the interpretation of the estimators should be different (Rousset, 1997). Nevertheless, the relevance of our method for estimating N_e could be a promising application, although its primary objective was estimating σ^2.

Chapitre 7

Conclusion

Dans la majeure partie de la thèse, nous nous sommes consacrés à l'étude de la forme des arbres phylogénétiques et plus particulièrement à leur déséquilibre. La structure d'arbre binaire intervient à la fois en phylogénie afin de représenter la généalogie des espèces et en génétique des populations afin de représenter la généalogie d'individus ou de gènes. Une généalogie est dite déséquilibrée si la plupart des noeuds internes de la généalogie (les ancêtres communs) séparent l'arbre en deux sous arbres de tailles sensiblement différentes. Le déséquilibre des arbres avait surtout été étudié en phylogénie. Les biologistes essayaient alors de retrouver quels étaient les modèles d'évolution susceptibles de reproduire des phylogénies ayant un déséquilibre comparable à celui des phylogénies connues.

D'un point de vue théorique, les propriétés des statistiques mesurant ce déséquilibre sont mal connues. Ces indices sont utilisés afin de détecter une phylogénie dont le déséquilibre est trop extréme par rapport aux prédictions d'un modèle nul. Le modèle canonique de généalogies de gènes en génétique des populations est le coalescent. En phylogénie, les généalogies d'espèces sont décrites par un processus de branchement. Le hasard mis en jeu dans ces deux modèles se situe à des échelles différentes. Il résulte de la stochasticité dans le nombre de descendants de chaque individu pour le premier tandis qu'il rend compte des fluctuations macroévolutives pour le second. D'un point de vue strictement mathématique, les deux modèles diffèrent puisque les temps de branchement ne suivent pas les mêmes lois que les temps de coalescence. Néanmoins, à nombre d'espèces total fixé, la topologie du

processus de branchement a exactement la même loi que celle du coalescent. Les variables aléatoires qui ne dépendent pas de la longueur des branches auront ainsi la même loi dans les deux modèles.

Nous avons obtenu les lois exactes d'un certain nombre de variables topologiques d'importance ainsi que les lois asymptotiques des statistiques de comptage correspondantes. La taille d'un clade aléatoire (le nombre de feuilles descendant d'un noeud tiré de manière uniforme parmi les noeuds internes) suit une loi puissance de paramètre 2. La taille d'un clade minimal (le nombre de plus proches parents d'un individu plus 1) suit une loi puissance de paramètre 3. Le nombre de clades (aussi appelé sous-arbres) de taille x ainsi que le nombre de clades minimaux de taille x sont asymptotiquement gaussiens. Les statistiques de comptage précédentes font partie d'une famille de statistiques dites additives. Les statistiques additives vérifient la relation de récurrence suivante :

$$S_n \stackrel{d}{=} S_{I_n} + S_{J_n} + t(I_n, J_n)$$

où $\stackrel{d}{=}$ représente l'égalité en loi et I_n (resp. J_n) le nombre de feuilles du sous arbre gauche (resp. droit) avec $I_n + J_n = n$. Dans le cas du modèle de Yule ou du coalescent, I_n suit une loi uniforme sur l'ensemble $\{1 \dots n-1\}$. Nous utilisons la terminolgie issue de l'analyse des algorithmes en appelant $t(I_n, J_n)$ la fonction coût. Une méthode de contraction récemment introduite pour l'analyse probabiliste des algorithmes (Rösler, 1991) nous a permis d'étudier l'asymptotique de ces variables additives.

Un certain nombre de statistiques utilisées pour détecter un écart au modèle nul de la phylogénèse sont, elles aussi, additives. Les deux statistiques les plus couramment utilisées sont les statistiques de Sackin et de Colless. La fonction coût de la statistique de Sackin est $t(I_n, J_n) = I_n + J_n = n$ et celle de l'indice de Colless est $t(I_n, J_n) = |I_n - J_n|$. La fonction côut correspondant aux nombres de clades est $t(I_n, J_n) = \delta_{I_n+J_n,x} = \delta_{n,x}$ et celle du nombre de clades minimaux est $t(I_n, J_n) = \delta_{n,x}(\delta_{I_n,1} + \delta_{I_n,n-1})$. La statistique de Fusco et Cronk (1995) modifiée par Purvis et al. (2002) ainsi que le premier indice B_1 de Shao et Sokal (1990) vérifient le même schéma récursif et pourraient être étudiés de la même manière.

Néanmoins, d'autres statistiques qui mesurent le déséquilibre d'une phylogénie

ne sont pas additives et leur distribution limite n'a pas été établie. De telles statistiques sont calculées à partir d'une somme sur les feuilles et non d'une somme sur les noeuds internes, et ne peuvent être exprimées sous forme récursive. C'est le cas de la deuxième statistique de Shao et Sokal (1990)

$$B_2 = \sum_{i=1}^{n} N_i / 2^{N_i}$$

où N_i est la profondeur d'une feuille i, i.e. le nombre de noeuds internes entre i et la racine (incluse). La deuxième statistique de Sackin σ_n^2 (1972) s'exprime aussi comme une somme sur les feuilles

$$\sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} (N_i - \bar{N})^2$$

où $\bar{N}$ est la moyenne empirique des N_i, i.e. la statistique de Sackin (divisée par n) dont on connaît la loi limite. La statistique de Sackin s'exprimait aussi comme une somme sur les feuilles mais du fait de sa linéarité par rapport aux N_i, elle pouvait s'exprimer de manière récursive. Les indices B_2 et σ_n^2 ne sont pas linéaires par rapport aux N_i ce qui rend leur étude difficile.

Nous nous sommes aussi intéressés aux lois de ces variables aléatoires dans le modèle uniforme qui suppose que les $(2n-3)!!$ phylogénies à n feuilles sont équiprobables. Souvent considéré par les biologistes comme dénué d'intérêt, ce modèle est en fait issu d'un processus évolutif. Aldous (1996) rappelle en effet, que la loi uniforme est la loi conditionnelle d'un arbre binaire de Galton-Watson sachant que le nombre total d'espèces est n (voir aussi Pinelis, 2003). Les combinatoriciens (Sedgewick et Flajolet, 1996) analysent les variables issues de modèles uniformes avec des méthodes analytiques reposant sur le calcul de fonctions génératrices. En utilisant leurs méthodes, nous avons pu établir les lois des tailles des clades.

- La taille d'un clade aléatoire suit une loi puissance de paramètre 3/2. Cette loi puissance décroît plus lentement que dans le modèle de Yule ($2 > 3/2$). C'est une manière de s'apercevoir que les arbres du modèle uniforme sont plus déséquilibrés.
- La taille d'un clade minimal suit aussi une loi puissance de paramètre 3/2

(Résultat facile à établir).

Dans le modèle de Yule et dans le modèle uniforme, nous retrouvons le résultat de Willis (1922) qui remarquait que le nombre d'espèces par genre suivait une loi puissance. L'apparition de la loi puissance est robuste quant au modèle de phylogénie considéré.

Une approche probabiliste permet aussi d'étudier les variables dans le modèle uniforme. Elle repose sur la correspondance entre un arbre et l'excursion discrète qui lui est associée. Aldous (1991a) a montré que cette excursion renormalisée converge vers l'excursion brownienne. Pour peu que la variable étudiée soit une fonction continue de l'excursion, la convergence en loi de la variable, celle de Sackin par exemple, résultera de la convergence des processus. Néanmoins, certaines variables ne sont pas des fonctions continues de cette excursion, par exemple l'indice de Colless, et cette méthode ne peut-être appliquée directement.

Nos travaux sur le déséquilibre des arbres étaient motivés par un phénomène précis : la transmission de la fertilité chez les humains. Nous avons montré que ce phénomène démographique déséquilibre fortement les généalogies de gènes. Les généalogies d'ADN mitochondrial de 41 populations humaines ont été reconstruites à partir de méthodes de reconstruction phylogénétique. Nos résultats montrent que les généalogies de gènes des populations de chasseurs-cueilleurs sont très déséquilibrées. Ceci suggère une forte transmission de la fertilité dans les populations de chasseurs-cueilleurs.

Certaines propriétes des généalogies, comme leur déséquilibre, sont porteuses d'information à la fois en phylogénie et en génétique des populations. Les ponts entre les deux disciplines sont multiples. Des modèles dédiés à priori à la génétique des populations peuvent être utilisés en phylogénie. Le modèle du chapitre 6 rend compte de l'evolution spatiale d'une population à partir de mouvements browniens. Les positions des individus échantillonnés ne peuvent être considérées comme indépendantes puisque les indidus partagent une généalogie commune, le coalescent. Nous avons proposé trois méthodes d'inférence du paramètre de dispersion spatiale défini comme étant la distance quadratique moyenne entre un individu et ses parents. Notons que l'évolution d'un caractère biologique partagé par plusieurs espèces a déja été modélisé par un mouvement brownien (Edwards,

1970 ; Felsenstein, 1985 ; Heard, 1992). Il a même été proposé d'utiliser un coalescent pour décrire la loi d'une phylogénie (Hey, 1992). Dans ce cas précis, la loi jointe des n traits biologiques est la même que la loi des n positions spatiales que nous avons étudiée. Les méthodes d'inférence du chapitre 6 permettent ainsi d'estimer les vitesses d'évolution de caractères biologiques intra-spécifiques ou inter-spécifiques.

Plus généralement, des indices de diversité phylogénétique viennent d'être introduits afin de mesurer la biodiversité (Mooers et al., 2005). Ces indices ressemblent aux mesures de diversité intra-spécifique utilisées en génétique des populations. Récemment, des probabilistes (Lambert, 2003 ; Popovitch, 2004) ont décrit les lois de la généalogie d'un processus de branchement. Les lois des mesures de biodiversité pourront ainsi être calculées dans le cadre de ce modèle. C'est une extension des travaux portant sur le déséquilibre des arbres puisque ces mesures de biodiversité prennent en compte l'aspect temporel. Elles saisiront de manière plus précise les différentes forces qui ont faconné la biodiversité.

7.1 English Conclusion

In the main part of this manuscript, we focused on the shape of phylogenetic trees and more specifically on their imbalance. Probability distributions of random variables that capture imbalance were poorly known. We studied here some of their properties in the Yule model which assumes that all species are equally likely to speciate.

We prove that under this model the probability distribution of the size of a random clade - the number of leaves descending from an internal node drawn uniformly amongst all the internal nodes - is a power law with parameter 2. The probability distribution of the size of the minimal clade-the number of closest relatives of an individual-is a power law distribution with parameter 3. Probability distributions of these random variables have also been established in the uniform model which assumes that all phylogenies with n leaves are equally likely. Both random variables have a power law distribution with parameter 3/2 (concerning the size of the minimal clade, it has not been proved in the present manuscript

but it is easy to obtain). Since 3/2 is lower than 2, subtrees containing a high number of leaves are more likely in the uniform model than in the Yule model. We may notice that the power law distribution arises in both models.

In order to measure the imbalance of phylogenetic trees, biologists introduced a wide number of one dimensional statistics. Many of them verify the following recurrence equation

$$S_n \stackrel{d}{=} S_{I_n} + S_{J_n} + t(I_n, J_n)$$

where $\stackrel{d}{=}$ denotes equivalence in distribution and I_n (resp. J_n) denotes the number of leaves in the left (resp. right) subtree ($I_n + J_n = n$). In the Yule model and in Kingman's coalescent, I_n is unifomly distributed on $\{1 \dots n-1\}$. Statistics verifying the above recurrence equation are called additive (Fill and Kapur, 2003). $t(I_n, J_n)$ is called the toll function. A contraction method introduced for the probabilistic analysis of algorithm (Rösler, 1991) is the clue for proving the weak convergence of additive statistics.

The Sackin's (1972) statistic is an additive statistic whose toll function is $t(I_n, J_n) = I_n + J_n = n$. The toll function of the Colless' (1982) statistic is $t(I_n, J_n) = |I_n - J_n|$. The number of subtrees of size x is also an additive statistic with a toll function equal to $t(I_n, J_n) = \delta_{I_n+J_n,x} = \delta_{n,x}$ and the toll function of the number of minimal clades is $t(I_n, J_n) = \delta_{n,x}(\delta_{I_n,1} + \delta_{I_n,n-1})$. The limiting distributions of the Colless' and the Sackin's statistic have also been derived in the uniform model. The proofs are based on the one to one correspondence between binary trees and random excursions (Aldous, 1991a).

In population genetics, the imbalance of gene genealogies appears powerful for detecting a specific departure from neutrality: fertility inheritance. We show that fertility inheritance unbalances gene genealogies even if the fertility was inherited only during a small number of generations. We have reconstructed the gene genealogies of 41 human populations from samples of mitochondrial DNA. Genealogies from hunter-gatherer populations are the most unbalanced, suggesting an high fertility inheritance in these populations.

Some features of genealogies such as their shape, contain information in population genetics and phylogeny. Links between these two fields are numerous. Models dedicated to population genetics can be used in phylogeny. The model

introduced in the chapter 6 describes the spatial evolution of a population using Brownian motions. The sampled individuals can not be considered independent because they share an unknown genealogy called coalescent. We propose three methods for estimating the spatial dispersal parameter. This parameter is defined as the expected quadratic distance between an individual and its parents. The evolution of biological characters has already been modelled by a Brownian motion (Edwards, 1970; Felsenstein, 1985; Heard, 1992). It has even be proposed to use the coalescent process for describing a phylogeny (Hey, 1992). In that case, the joint distribution of the n biological characters is the same as the joint distribution of the n spatial locations. The inference methods introduced in the chapter 6 can be used both in an intra-specific and an inter-specific framework.

More generally, indices of phylogenetic diversity are introduced in order to measure biodiversity (Mooers, 2005). These indices have a lot in common with classical indices of genetic diversity used in population genetics. Recently probabilists (Lambert, 2003; Popovitch, 2004) studied the genealogy of extant species in a branching process framework. Using their results, it will be possible to compute the probability distribution of phylogenetic indices that measure biodiversity. It is an extension of the work on tree balance since these measures take the branch length information into account. Phylogenetic diversity will capture more precisely various phenomenae which have structured biodiversity.

Bibliography

[1] Abramowitz M. and I. Stegun 1970. *Handbook of mathematical functions.* Dover, New York.

[2] Agapow P-M. and A. Purvis 2002. Power of eight tree shape statistics to detect nonrandom diversification: a comparison by simulation of two models of cladogenesis. *Syst. Biol.* **51**(6) 866-872.

[3] Aldous D.J. 1991a. The continuum random tree II. In: *Stochastic analysis*, N.T. Barlow and N.H. Bingham eds. Cambridge University Press 23-70.

[4] Aldous, D.J. 1991b. Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.* 1 228-266.

[5] Aldous D.J. 1995. Probability Distributions on Cladograms. In: *Random Discrete Structures*, D. Aldous and R. Pemantle eds. Springer Berlin 1-18.

[6] Aldous D.J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science* **16** 23-34.

[7] Athreya K.B. and P.E. Ney 1972. *Branching Processes.* Springer, Berlin

[8] Austerlitz F. and E. Heyer 1998. Social transmission of reproductive behavior increases frequency of inherited disorders in a youg-expanding population. *Proc. Natl. Acad. Sci. USA* **95** 15140-15144.

[9] Austerlitz F. and E. Heyer 2000. Allelic association is increased by correlation of effective family size. *Eur. J. Hum. Genet.* **8** 980-985.

[10] Austerlitz F., L. Kalaydjieva and E. Heyer 2003. Detecting Population Growth, Selection and Inherited Fertility From Haplotypic Data in Humans. *Genetics* **165** 1579-1586.

[11] Barton N.H. 1998. The effect of hitch-hiking on neutral genealogies. *Genet. Res. Camb.* **72** 123-133.

[12] Barton N.H., F. Depaulis and A.M. Etheridge 2002. Neutral evolution in spatially continuous populations. *Theor. Popul. Biol.* **61** 31-48.

[13] Beaumont, M.A. 2003. Conservation genetics. In *Handbook of Statistical Genetics*. D.J. Balding, M.J. Bishop, and C. Cannings eds. John wiley & Sons, Chichester, UK, 779-812.

[14] Beerli P. and J. Felsenstein 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. USA* **98** 4563-4568.

[15] Beerli P. 2002. MIGRATE: documentation and program, part of LAMARC. Version 1.5. Revised August 7, 2002. Distributed over the Internet, http://evolution.genetics.washington.edu/lamarc.html.

[16] Blum M.G.B. and O. François 2005a. On statistical tests of phylogenetic imbalance: the Sackin and other indices revisited. *Math. Biosci.* **195** 141-153.

[17] Blum M.G.B. and O. François 2005b. Minimal clade size and external branch length under the neutral coalescent. *Adv. in Appl. Probab.* **37** (3) 647-662.

[18] Bossart J.L. and D.P. Prowell 1998. Genetic estimates of population structure and gene flow: Limitations, lessons, and new directions. *Trends Ecol. Evol.* **13** 171-212.

[19] Brown J.K.M. 1994. Probabilities of evolutionary trees. *Syst. biol.* **43** 78-91.

[20] Caballero A. 1994. Developments in the prediction of effective population size. *Heredity* **73** 657-679.

[21] Campbell R.B. 1999. The coalescent time in the presence of background fertility selection. *Theor. Pop. Biol.* **55** 260-269.

[22] Campbell K.L. and J.W Wood 1988. Fertility in traditional societies: Social and Biological Determinants. In *Natural Human Fertility*. P. Diggory, M. Potts, and S. Teper eds. Macmillan, Hampshire, U.K., 39-69.

[23] Cann R.L., M. Stoneking and A.C. Wilson 1987. Mitochondrial DNA and human evolution. *Nature* **325** 31-36.

[24] Cavalli-Sforza L.L. and M.W. Feldman 1991. *Cultural transmission and evolution: a quantitative approach*. Princeton University Press.

[25] Chagnon N.A. 1979. Is reproductive success equal in egalitarian societies? In *Evolutionary Biology and Human Social Behavior*. N.A. Chagnon and W. Irons eds. North Scituate: Duxbury, 374-401.

[26] Colless D.H. 1982. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Syst. Zool.* **31** 100-104.

[27] Cox J.T. and R. Durrett 2002. The stepping stone model: New Formulas expose old myths. *Ann. Appl. Probab.* **12** 1348-1377

[28] Darwin C. 1859. *The origin of species*. Reprinted by Penguin books, London, UK.

[29] Devroye L. 1991. Limit laws for local counters in random binary search trees. *Random Structures Algorithms* **2** 303-315.

[30] Di Rienzo A. and A.C. Wilson 1991. Branching Pattern in the Evolutionary Tree for Human Mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **88** 1597-1601.

[31] Donnelly P., S. Tavaré, D.J. Balding and R.C. Griffiths 1996. Estimating the age of the common ancestor of men from the ZFY intron. *Science* **272** 1357-1359.

[32] Draper P. and R. Hames 2000. Birth order, sibling investment, and fertility among Ju/'hoansi (!Kung) *Human nature-an interdisciplinary biological perspective* **11** 117-156.

[33] Dumond D.E 1975. Limitation of human population - a natural history. *Science* **187** (4178): 713-721.

[34] Durrett R. 2003. *Probabilistic models of DNA sequences.* Springer-Verlag, New-York.

[35] Edwards A.W.F. 1970. Estimation of the branch points of a branching diffusion process. *J. R. Statist. Soc. B* **32** 155-174.

[36] Ethier S.N. and T.G. Kurtz 1986. *Markov processes, Characterization and Convergence.* Wiley, New York.

[37] Excoffier L. and A. Langaney 1989. Origin and differentiation of human mitochondrial data. *Am. J. Hum. Genet.* **44** 73-85.

[38] Excoffier L. and S. Schneider 1999. Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. *Proc. Natl. Acad. Sci. USA* **96** 10597-10602.

[39] Excoffier L., J. Novembre and S. Schneider 2000. Computer note. SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J. Hered.* **91** 506-509.

[40] Excoffier L. 2004. Patterns of DNA sequence diversity and genetic structure after a range expansion: lessons from the infinite-island model. *Mol. Ecol.* **13** 853-864.

[41] Fay J.C. and C.I. Wu 2000. Hitchiking under positive darwinian selection. *Genetics* **155** 1405-1413.

[42] Felsenstein J. 1971. The rate of loss of multiple alleles in finite haploid populations. *Theor. Pop. Biol.* **2** 391-403.

[43] Felsenstein J. 1975. A pain in the torus: Some difficulties with models of isolation by distance. *Am. Nat.* **109** 359-368.

[44] Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach *J. Mol. Evol.* **17** 368-376.

[45] Felsenstein J. 1985. Phylogenies and the comparative method. *Am. Nat.* **125** 1-15.

[46] Felsenstein J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5** 164-166.

[47] Felsenstein J., M.K. Kuhner, J. Yamato and P. Beerli 1999. Likelihoods on coalescents: a Monte Carlo Sampling approach to inferring parameters from population samples of molecular data. Statistics in Molecular Biology, IMS lecture notes Monograph Series **33** 163-185.

[48] Fill J.A. and N. Kapur 2004. Limiting distributions for additive functionals on Catalan trees. *Theoretical Computer Science* 326 69-102.

[49] Fisher R.A. 1930. *The Genetical Theory of Natural Selection.* Oxford University Press.

[50] Flajolet P. and G. Louchard 2001. Analytic Variations on the Airy Distribution. *Algorithmica* **31** 361-377.

[51] Fu Y.X. and W.H. Li 1993a. Statistical tests of neutrality of mutations. *Genetics* **133** 693-709.

[52] Fu Y.X. and W.H. Li 1993b. Maximum likelihood estimation of population parameters, *Genetics* **134** 1261-1270.

[53] Fusco G. and Q.C.B. Cronk 1995. A new method for evaluating the shape of large phylogenies. *J. theor. Biol.* **175** 235-243.

[54] Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14** 685-695.

[55] Gascuel O. 2000. Evidence for a relationship between algorithmic scheme and shape of inferred trees in *Data Analysis, Scientific Modeling and Practical Applications*, W. Gaul, O. Opitz and M. Schader eds. Springer, Berlin, 157-168.

[56] Goldstein D.B., A.R. Linares, L.L. Cavalli-Sforza and M.W. Feldman. 1995. An evaluation of genetic distances for use microsatellite loci, *Genetics* **139** 463-471.

[57] Goody E. 2005. 'Stone Age sociology' - Or sociology of early language-using society? *Journal of the royal anthropological institute.* **11** 585-588.

[58] Gould S.J., D.M. Raup , J.J. Sepkoski, T.J.M. Schopf and D.S. Simberloff 1977. The shape of evolution: a comparison of real and random clades. *Paleobiology* **3** 23-40.

[59] Griffiths R.C. 1980. On the distribution of allele frequencies in a diffusion model. *Theor. Pop. Biol.* **15** 140-158.

[60] Griffiths R.C. and S. Tavaré 1994a. Ancestral inference in population genetics. *Stat. Sci.* **9** 307-319.

[61] Griffiths R.C. and S. Tavaré 1994b. Simulating probability distributions in the coalescent. *Theor. Pop. Biol.* **46** (2) 131-159.

[62] Griffiths R.C. 2003. The frequency spectrum of a mutation, and its age, in a general diffusion model. *Theor. Pop. Biol.* **64** 241-251.

[63] Guglielmino C.R., C. Viganotti, B. Hewlett and L.L. Cavalli-Sforza 1995. Cultural Variation in Africa: Role of Mechanisms of Transmission and Adaptation *Proc. Natl. Acad. Sci. USA* **92** 7585-7589.

[64] Guindon S. and O. Gascuel 2003. A Simple, Fast and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.* **52** 696-704.

[65] Guyer C. and J.B. Slowinski 1991. Comparisons between observed phylogenetic phylogenies with null expectations among three monophyletic lineages. *Evolution* **45** 340-350.

[66] Guyer C. and J.B. Slowinski 1993. Adaptative radiation and the topology of large phylogenies. *Evolution* **47** 253-263.

[67] Handwerker W.P. 1983. The 1st demographic transition - an analysis of subsitence choices and reproductive consequences. *American anthropologist* **85** 5-27.

[68] Harding E.F. 1971. The probabilities of rooted tree shapes generated by random bifurcation. *Adv. in Appl. Probab.* **3** 44-77.

[69] Harvey P.H., A.J. Leigh Brown, J. Meynard Smith and S. Nee 1996. *New uses for new phylogenies.* Oxford University Press.

[70] Hasegawa M., H. Kishino and T. Yano 1985. Dating the human-ape splitting by amolecular clock of mitochondrial DNA. *J. Mol. Evol.* **22** 160-174.

[71] Heard S.B. 1992. Patterns in tree balance among cladistic, phenetic and randomly generated phylogenetic trees. *Evolution* **46** 1818-1826.

[72] Heard S.B. 1996. Patterns in phylogenetic tree balance with variable and evolving speciation rate. *Evolution* **50** 2141-2148.

[73] Hedrick P.W. 2003. Hopi Indians, "cultural" selection and albinism. *Am. J. Phys. Anthropol.* **121** 151-156.

[74] Helgason A., B. Hrafnkelsson, J.R. Gulcher, R. Ward and K. Stefansson 2003. A Populationwide Coalescent Analysis of Icelandic Matrilineal and Patrilineal Genealogies: Evidence for a faster Evolutionaary Rate of mtDNA Lineages than Y Chromosomes. *Am. J. Hum. Genet.* **72** 1370-1388.

[75] Hennig W. 1966. *Phylogenetic Systematics*. University Illinois Press, Urbana.

[76] Hey J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution **46** 627-640.

[77] Heyer E., Sibert A. and F. Austerlitz. 2005. Cultural transmission of fertility: genes takes the fast lane. *Trends in Genetics*. In press.

[78] Hill K. and A.M. Hurtado 1996. *Ache Life History: The Ecology and Demography of a Foraging People*. Aldine de Gruyter.

[79] Hoare C.A.R. 1962. Quicksort. *Computer Journal* **5** 10-15.

[80] Huelsenbeck J.P. and M. Kirkpatrick 1996. Do phylogenetic methods produce trees with biased shapes? *Evolution* **50** 1418-1424.

[81] Hwang H-K. and R. Neininger 2002. Phase change of limit laws in the quicksort recurrence under varying toll functions. *SIAM J. Comput.* **31** 1687-1722.

[82] Janson S. 2003. The Wiener index of simply generated random trees. *Random Structures and Algorithms* **22** 337-358.

[83] Kimura M. 1953. Stepping-stone model of population *Annual Report of the National Institute of Genetics*, Japan, **3** 62-63

[84] Kingman J.F.C. 1982a. The coalescent, *Stoch. Proc. Appl.* **13** 235-248.

[85] Kingman J.F.C. 1982b. On the genealogies of large populations. *Journal of Applied Probability* **19A** 27-43.

[86] Kirkpatrick M. and M. Slatkin. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree, *Evolution* **47** 1171-1181.

[87] Klein E.K., F. Austerlitz and C. Larédo 1999. Some statistical improvements for estimating population size and mutation rate from segregating sites in DNA sequences. *Theor. Popul. Biol.* **55** 235-247.

[88] Knuth D.E. 1973. *The Art of Computer Programming*, Volume 3: Sorting and Searching. Addison-Wesley, Reading, MA.

[89] Korber B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, R. Lapedes, B. Hahn, S. Wolinsky and T. Battacharya 2000. Timing the ancestor of the HIV-1 pandemic strains, *Science* **288** 1789-1796.

[90] Kring M., A. Stone, R.W. Schmitz, H. Krainitzki, M. Stoneking and S. Paabo 1997. Neandertal DNA sequences and the origin of modern humans. Cell **90** 19-30.

[91] Kuhner M.K., J. Yamato and J. Felsenstein 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling *Genetics* **140** 1421-1430.

[92] Lambert A. 2003. Coalescence times for the branching process. *Adv. Appl. Probab.* **35** 1071-1089.

[93] Mace R. 1996. Biased parental investment and reproductive success in Gabbra pastoralists. *Behavioral ecology and sociobiology* **38** 75-81.

[94] Mahmoud H. 1992. *Evolution of Random Search Trees*. Wiley, New York.

[95] Malécot G. 1967. Identical loci and relationship, *Proc. Fifth Berkeley Sy mp. Math. Stat. Prob.* **4** 317-332, Univ. California Press, Berkeley.

[96] Maia L.P., A. Colato and J.F. Fontanari 2004. Effect of selection on the topology of genealogical trees. *J. Theor. Biol.* 226 315-320.

[97] Martinez C., A. Panholzer and P. Helmut 1998. The number of descendants and ascendants in random search trees. *Electronic Journal of Combinatorics* **5**.

[98] McKenzie A. and M.A. Steel 2000. Distributions of cherries for two models of trees. *Math. Biosci.* **164** 81-92.

[99] McKenzie A. and M.A. Steel 2001. Properties of phylogenetic trees generated by Yule-type speciation models, *Math. Biosci.* **170** 91-112.

[100] Millstein R.L. 2000. Chance and macroevolution. *Philosophy of Science* **67** 603-24.

[101] Mooers A.O. 1995. Tree balance and tree completeness. *Evolution* **49** 379-384.

[102] Mooers A.O. and S.B. Heard 1997. Inferring evolutionary process from phylogenetic tree shape. *Quart. Rev. Biol.* **72** 31-54.

[103] Mooers A.O., S.B. Heard and E. Chrostowski 2005. Evolutionary heritage as a metric for conservation. in *Phylogeny and Conservation*, A. Purvis, T.L. Brooks and J.L. Gittleman, eds. Oxford University Press, Oxford.

[104] Mulder M.B. 1998. Brothers and sisters - How sibling interactions affect optimal parental allocations. *Human nature - an interdisciplinary biological perspective* **9** 119-161.

[105] Murphy M. and D.L. Wang 2001. Family-level continuities in childbearing in low-fertility societies. *European journal of population* **17** 75-96.

[106] Murray-McIntosh R.P., B.J. Scrimshaw, P.J. Hatfield and D. Penny 1998. Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. *Proc. Natl. Acad. Sci. USA* **95** 9047-9052.

[107] Nagylaki T. 2002. When and where was the most recent common ancestor? *J. math. Biol.* **44** 253-275.

[108] Neel J.V. 1970. Lessons from a "primitive" people. *Science* **170** 815-822.

[109] Neel J.V. and K.M Weiss 1975. Genetic structure of a tribal population, Yanomana indians. Biodemographic studies. *American Journal of physical anthropology* **42** 25-51.

[110] Nei M. 1987. *Molecular Evolutionary Genetics*, Columbia University Press, New York.

[111] Nei M. and M. Murata 1966. Effective population size when fertility is inherited. *Genet. Res.* **8** 257-260.

[112] Nei M. and S. Kumar 2000. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York.

[113] Neininger R. 2001. On a multivariate contraction method for random recursive structures with applications to quicksort. *Random Structures and Algorithms* **19** 498-524.

[114] Neininger R. 2002. The Wiener index of random trees. *Combinatorics, Probability and Computing* **11** 587-597.

[115] Nielsen R. 2001. Distinguishing migration from isolation: an MCMC Approach. *Genetics* **158** 885-896.

[116] Nordborg M. 1998. On the Probability of Neanderthal Ancestry. *Am. J. Hum. Genet.* **63** 1237-1240.

[117] Nordborg M. 2001. Coalescent theory. In *Handbook of statistical genetics*, D.J. Balding et al eds, Wiley and sons, inc, New-York, 179-208.

[118] Otha T. and M. Kimura 1972. On the stochastic model for estimation of mutational distance between homologous proteins, *J. Mol. Evol.* **2** 87-90.

[119] Otter R. 1949. The multiplicative process. *Annals of Mathematical Statistics*, Baltimore, Md. **20** 206-224.

[120] O'Ryan C., M. Bruford, W.M. Beaumont, R.K. Wayne, M.I. Cherry and E.H. Harley 1998. Genetics of fragmented populations of African buffalo (Syncerus caffer) in South Africa, *Animal Conservation* **1** 85-94.

[121] Paetkau D., L.P. Waits, L. Craighead, P. Clarkson and C. Strobeck 1998. Dramatic variation in genetic diversity across the range of North America brown bears, *Conserv. Biol.* **12** 418-429.

[122] Page R.D.M. 1990. Component analysis: a valiant failure? *Cladistics* **6** 119-136.

[123] Panchen A.L. 1992. *Evolution, Classification and the Nature of Biology.* Cambridge: Cambridge University Press.

[124] Paradis E., J. Claude and K. Strimmer 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20** 289-90.

[125] Parsons T.J., D.S. Muniec, K. Sullivan, N. Woodyatt, R. Alliston-Greiner, M.R. Wilson, D.L. Berry, K.A. Holland, V.W. Weedn, P. Gill and M.M. Holland 1997. A high observed substitution rate in the human mitochondrial DNA control region. *Nat. Genet.* **15** 363-368.

[126] Peterson J.A. 1983. The evolution of the subdigital pad in Anolis.I.Comparisons among the anoline genera, in A.G.J. Rhodin and K. Miyata eds. *Advances in Herpetology and evolutionary biology: essays in honor of Ernest E. Williams.* Cambridge, MA: Museum of comparative zoology, Harvard University, 245-283.

[127] Pinelis I. 2003. Evolutionary models of phylogenetic trees. *Proc. R. Soc. Lond. B* **270** 1425-1431.

[128] Popovitch L. 2003. Asymptotic Genealogy of a Critical Branching Process. *Ann. Appl. Probab.* **14** 2120-2148.

[129] Przeworski M., B. Charlesworth and J.D. Wall 1999. Genealogies and weak purifying selection. *Mol. Biol. Evol.* **16** 246-252.

[130] Pritchard J. K. and M.W. Feldman 1996. Statistics for Microsatellite Variation Based on Coalescence, *Theor. Popul. Biol.* **50** 325-344.

[131] Purvis A., A. Katzourakis and P.M. Agapow 2002. Evaluating Phylogenetic Tree Shape: Two modifications to Fusci and Cronk's Method. *J. Theor. Biol.* **214** 99-103.

[132] Pybus O.G., A. Rambaut and P.H. Harvey 2000. An Integrated Framework for the Inference of Viral Popultion History From Reconstructed Genealogies. *Genetics* **155** 1429-1437.

[133] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2003.

[134] Rachev S.T. and L. Rüschendorf 1995. Probability metrics and recursive algorithms. *Advances in Applied Probability* **27** 770-799.

[135] Rambaut A. and N.C. Grassly 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13** 235-238.

[136] Rambaut A., D.L. Robertson, O.G. Pybus, M. Peeters and E.C Holmes. Human immunodeficiency virusPhylogeny and the origin of HIV-1. *Nature* 2001 **410** 1047-1048.

[137] Raup D.M., S.J. Gould, T.J.M. Schopf and D.S. Simberloff 1973. Stochastic models of phylogeny and the evolution of diversity. *Journal of Geology* **81** 525-542.

[138] Ray N., M. Currat and L. Excoffier 2003. Intra-deme molecular diversity in spatially expanding populations *Molecular biology and evolution* **20** 76-86.

[139] Raymond M. and F. Rousset 1995. GENEPOP: a population genetics software for exact tests and ecumenicism, *J. Hered.* **86** 248-249.

[140] Regnier M. 1989. A limiting distribution for quicksort. *RAIRO Information Theor. Appl.* **23** 335-343.

[141] Robert C.P. and G. Casella 1999. *Monte Carlo Statistical Methods* Springer Series in Statistics, Springer-Verlag, New York.

[142] Rogers J.S. 1993. Response of Colless's tree imbalance to number of terminal taxa. *Systematic Biology* **42** 102-105.

[143] Rogers J.S. 1994. Central moments and probability distribution of Colless's coefficient of tree imbalance. *Evolution* **48** 2026-2036.

[144] Rogers J.S. 1996. Central moments and probability distrbution of three measures of phylogenetic tree imbalance. *Syst. Biol.* **45** 99-110.

[145] Rosenberg N. 2004. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Preprint submitted.

[146] Rösler U. 1991. A limit theorem for quicksort. *RAIRO Inform. Theor. Appl.* **25** 85-100.

[147] Rösler U. 2001. The analysis of divide and conquer algorithms. *Algorithmica* **29** 238-261.

[148] Roth E.A. 1985. A note on the demographic concomitants of sedentism. *American anthropologist* **87** 380-382.

[149] Rousset F. 1997. Genetic differentiation and estimation of gene flow using F-statistics under isolation by distance,*Genetics* **145** 1219-1228.

[150] Rousset F. 2003. Inferences from spatial population genetics, In *Handbook of Statistical Genetics* , D. J. Balding, M. J. Bishop, and C. Cannings, eds., John wiley & Sons, Chichester, UK, 239-270.

[151] Runciman W.G. 2005. Stone age sociology. *Journal of the royal anthropological institute* **11** 129-142.

[152] Sackin M.J. 1972. "Good" and "bad" phenograms. *Systematic Zoology* **21** 225-226.

[153] Saunders I.W., S. Tavaré and G.A. Watterson 1984. On the genealogy of nested subsamples from a haploid population. *Adv. in Appl. Probab.* **16** 471-491.

[154] Savage H.M. 1983. The shape of evolution: systematic tree topology. *Biol. J. Linnean Soc.* **20** 225-244.

[155] Schopf T.J.M. 1979. Evolving paleontological views on determinism and stochastic approaches. *Paleobiology* **5** 337-352.

[156] Sedgewick R. and P. Flajolet 1996. *An Introduction to the Analysis of Algorithms.* Addison Wesley.

[157] Seo T.K., J.L. Thorne, M. Hasegawa and H. Kishino 2002. Estimation of Effective Population Size of HIV-1 Within a Host: A Pseudomaximum-Likelihood Approach, *Genetics* **160** 1283-1293.

[158] Shao K. and R.R. Sokal 1990. Tree balance. *Systematic Zoology* **39** 266-276.

[159] Slatkin M. 1993. Isolation by distance in equilibrium and non-equilibrium populations. *Evolution* **47** 264-279.

[160] Sibert A., F. Austerlitz and E. Heyer 2002. Wright-fisher revisited: the case of fertility correlation. *Theor. Pop. Biol.* **62** 181-197.

[161] Simberloff D., K.L. Heck, E.D. McKoy and E.F. Connor 1981. There have been no statistical tests of cladistic biogeographical hypotheses. In G. Nelson, D.E. Rosen, editors. *Vicariance biogeography: A critique.* New York: Columbia University Press 40-63.

[162] Slatkin M. and R.R. Hudson 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* **129** 555-562.

[163] Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139** 457-462.

[164] Stephens M. 2003. Inference under the coalescent, In *Handbook of Statistical Genetics* D.J. Balding, M.J. Bishop, and C. Cannings, eds., John wiley & Sons, Chichester, UK, 213-238.

[165] Taberlet P. and and J. Bouvet 1994. Mitochondrial DNA polymorphism, phylogeography, and conservation genetics of the brown bear (Ursus arctos) in Europe., *P. Roy. Soc. Lond. B Bio.* **255** 195-200.

[166] Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105** 437-460.

[167] Takacs L. 1991. A Bernoulli excursion and its various applications. *Adv. Appl. Probab.* **23** 557-585.

[168] Tamura K. and M. Nei 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Bio. Evol.* **10** 512-526.

[169] Tan H.K. and P. Hadjicostas 1995. Some properties of a limiting distributions of quicksort. *Statist. Probab. Lett.* **25** 87-94.

[170] Tavaré S. Ancestral inference from DNA sequence data. In *Mathematical modeling in Ecology, Physiology and cell biology*, chap 5, Othmer at al eds, 1997, 81-96.

[171] Tavaré S. 2003. *Lecture notes St Flour*. Springer-Verlag.

[172] Vidal N., M. Peeters, C. Mulanga-Kayeba, N. Nzilambi, D. Robertson, W. Ilunga, H. Sema, K. Tshimanga, B. Bongo and E. Delaporte 2000. Unprecedented degree of HIV-1 group M genetic diversity in the Democratic Republic of Congo suggests that the HIV-1 pandemic originated in Central Africa. *J. Virol.* **74** 498-507.

[173] Waits L., P. Taberlet, J.E. Swenson, F. Sandegren and R. Franzen 2000. Nuclear DNA microsatellite analysis of genetic diversity and gene flow in the Scandinavian brown bear (Ursus arctos). *Mol. Evol.* **9** 610-621.

[174] Wakeley J. 2005. The Limits of Theoretical Population Genetics. *Genetics* **169** 1-7.

[175] Walsh B. 2001. Estimating the time to the most recent common ancestor for the Y chromosome or mitochontrial DNA for a pair of individuals, *Genetics* **158** 897-912.

[176] Watterson G.A. 1975. On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* **7** 256-276.

[177] Watterson G.A. 1982. Mutant substitutions at linked nucleotide sites. *Adv. Appl. Probab.* **14** 206-224.

[178] Whitehead H. 1998. Cultural selection and genetic diversity in matrilineal whales. *Science* **282** 1708-1711.

[179] Williams L.A. and B.J Williams 1974. A re-examination of the heritability of fertility in the British peerage. *Soc. Biol.* **21** 225-231.

[180] Willis J.C. 1922. *Age and Area.* Cambridge: Cambridge university press.

[181] Wilson I.J. and D.J. Balding 1998. Genealogical inference from microsatellite data, *Genetics* **150** 499-510.

[182] Wiuf C. and P. Donnelly. 1999. Conditional genalogies and the age of a neutral mutant, *Theor. Pop. Biol.* **56** 183-201.

[183] Wright S. 1931. Evolution in Mendelian populations. *Genetics* **16** 97-159

[184] Wright S. 1943. Isolation by distance. *Genetics* **28** 114-138.

[185] Yule G.U. 1924. A mathematical theory of evolution, based on the conclusions of Dr J.C. Willis. *Philos. Trans. Roy. Soc. London Ser. B* **213** 21-87.

[186] Yusim K., M. Peeters, O.G. Phybus, T. Bhattacharya, E. Delaporte, C. Mulanga, M. Muldoon, J. Theiler and B. Korber 2001. Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. *Phil. Trans. R. Soc. Lond. B* **356** 855-866.

[187] Zerjal T., Y. Xue, G. Bertorelle, R.S. Wells, W. Bao, S. Zhu, R. Qamar, Q. Ayub, A. Mohyuddin, S. Fu, P. Li, N. Yuldasheva, R. Ruzibakiev, J. Xu, Q. Shu, R. Du, H. Yang, M.E. Hurles, E. Robinson, T. Gerelsaikhan, B. Dashnyam, Q. Mehdi and C. Tyler-Smith 2003. The genetic legacy of the Mongols. *Am. J. Hum. Genet.* **72** 717-721.

Appendix A

Convergence of compound Poisson process toward Brownian motion

This appendix is related to the chapter 5.

Let us prove that Brownian motion can be obtained as limits of compound Poisson process which includes the stepwise mutation model.

Let $(\xi_i)_{i\geq 1}$ be independent and identically distributed random variables such that

$$E[\xi_i] = 0 \qquad \text{and} \qquad Var[\xi_i] = \nu^2, \quad \nu > 0.$$

Let $\theta > 0$. Consider a family of homogenous Poisson process (M_t^p) of rate θp $(p \geq 1)$ where M_t^p is the number of occurrences at time t. Define the *compound Poisson process* (X_t^p) as follows

$$X_t^p = \frac{1}{\nu\sqrt{p}} \sum_{i=1}^{M_t^p} \xi_i \tag{A.1}$$

Consider Bernoulli random variables

$$P(\xi_i = 1) = P(\xi_i = -1) = 1/2.$$

In this situation, we have $\nu^2 = 1$. For $p = 1$, M_t^1 corresponds to the number of mutations at time t in the stepwise mutation model, and X_t^1 is the number of

differences between the ancestral allelic state and the current state.

We consider more general mutation models than the ladder model, and establish the weak convergence of X_t^p to B_t.

Theorem 1 *Let $(X_\cdot^p)$ be the stochastic process defined in equation A.1 and $(B_\cdot)$ be a one-dimensional standard Brownian motion times $\sqrt{\theta}$. We have*

$$X_\cdot^p \stackrel{\mathcal{D}}{\Rightarrow} B_\cdot, \quad \text{as } p \to \infty.$$

where $\mathcal{D}$ denotes the weak convergence in $\mathcal{D}_{\mathcal{R}}(0,\infty)$, the set of càd-làg functions defined on $(0,\infty)$.

Proof. For all $t > 0$ and $p \geq 1$, the first and second moments of X_t^p are

$$E\left[X_t^p\right] = 0$$

and

$$\begin{aligned} Var\left[X_t^p\right] &= E\left[(X_t^p)^2\right] = E\left[E\left[(X_t^p)^2 \mid M_t^p\right]\right] \\ &= \frac{1}{\nu^2 p} E[E[\sum_{i=1}^{M_t^p} \xi_i^2 \mid M_t^p]] = \frac{1}{\nu^2 p} E\left[M_t^p \nu^2\right] \\ &= \theta t. \end{aligned}$$

According to the Theorem 7.8 in Chapter 3 of (36), the result follows from the convergence of the finite dimensional distributions and the relative compactness of $(X_\cdot^p)$ which are demonstrated below. ■

Convergence of the finite-dimensional distributions

Let

$$Y_t^p = \frac{1}{\nu\sqrt{p}} \sum_{i=1}^{\lfloor \theta p t \rfloor} \xi_i. \tag{A.2}$$

First, we show that the random variables X_t^p converge weakly toward B_t for fixed $t > 0$

$$X_t^p \stackrel{\mathcal{D}}{\Rightarrow} B_t.$$

Lemma 5 *Let (X_t^p) and (Y_t^p) be defined in equations A.1 and A.2. Then, $(Y_t^p - X_t^p)$ converges in probability to 0 as $p \to \infty$*

Proof. Let $p \geq 1$ and $\varepsilon > 0$. By the stationarity of the ξ_i's, we have

$$\begin{aligned} P(|X_t^p - Y_t^p| \geq \varepsilon) &= P(\frac{1}{\nu\sqrt{p}}|\sum_{i=\lfloor \theta pt \rfloor + 1}^{M_t^p} \xi_i| \geq \varepsilon) \\ &= P(\frac{1}{\nu\sqrt{p}}|\sum_{i=1}^{|M_t^p - \lfloor pt \rfloor|} \xi_i| \geq \varepsilon) \end{aligned}$$

Conditioning on M_t^p, this probability is equal to

$$\begin{aligned} &\sum_{n=0}^{\infty} P(|\sum_{i=1}^{|n-\lfloor \theta pt \rfloor|} \xi_i| \geq \nu\sqrt{p}\,\varepsilon) P(M_t^p = n) \\ &\leq \sum_{n=0}^{\infty} \frac{1}{\nu^2 p\,\varepsilon^2} Var[\sum_{i=1}^{|n-\lfloor \theta pt \rfloor|} \xi_i) P(M_t^p = n) \end{aligned}$$

and the upper bound follows from Chebyshev's inequality. Then we have

$$\begin{aligned} P(|X_t^p - Y_t^p| \geq \varepsilon) &\leq \sum_{n=0}^{+\infty} \frac{|n - \lfloor \theta pt \rfloor|}{p\varepsilon^2} P(M_t^p = n) \\ &\leq \frac{E[|M_t^p - \lfloor \theta pt \rfloor|]}{p\varepsilon^2} \\ &\leq \frac{E[|(M_t^p - \theta pt) + (\theta pt - \lfloor \theta pt \rfloor)|]}{p\varepsilon^2} \\ &\leq \frac{E[|M_t^p - \theta pt|] + E[|\theta pt - \lfloor \theta pt \rfloor|]}{p\varepsilon^2} \end{aligned}$$

where we use the triangle inequality. We finish with the Cauchy-Schwarz inequality

$$\begin{aligned} P(|X_t^p - Y_t^p| \geq \varepsilon) &\leq \frac{E[|M_t^p - \theta pt|] + 1}{p\varepsilon^2} \\ &\leq \frac{\sqrt{E[(M_t^p - \theta pt)^2]} + 1}{p\varepsilon^2} \end{aligned}$$

$$\leq \frac{\sqrt{Var[M_t^p]}+1}{p\varepsilon^2}$$
$$\leq \frac{\sqrt{\theta pt}+1}{p\varepsilon^2}$$

Since

$$\lim_{p\to\infty} \frac{\sqrt{\theta pt}+1}{p\varepsilon^2} = 0,$$

the convergence is established. ■

According to the central limit theorem and to the fact that $\lim_{p\to+\infty} \frac{\lfloor p\theta t \rfloor}{\theta p} = t$, we have

$$Y_t^p \xrightarrow{\mathcal{D}} B_t$$

By Lemma 5, Slutsky's Theorem (Eithier and Kurtz, 1986) ensures the convergence in law of X_t^p to B_t. Now, fix $t_n > \ldots > t_1 > 0$. Showing

$$(X_{t_1}^p, \ldots, X_{t_n}^p) \overset{\mathcal{D}}{\Rightarrow} (B_{t_1}, \ldots, B_{t_n})$$

amounts to prove that

$$(X_{t_1}^p, \ldots, X_{t_n}^p - X_{t_{n-1}}^p) \overset{\mathcal{D}}{\Rightarrow} (B_{t_1}, \ldots, B_{t_n} - B_{t_{n-1}})$$

This result can be easily checked from the independence of increments in the compound Poisson process. ■

Relative compactness of $(X_\cdot^p)_{p\geq 1}$

Let $\mathcal{K}$ be the set of compact subsets of $\mathcal{R}$. From Theorem 8.6 and 8.8 in chapter 3 of Ethier and Kurtz (1986), the following three conditions imply the relative compactness of $(X_\cdot^p)$.

- Condition (i)

$$\forall \eta > 0, \forall t \in \mathcal{Q}_+^*, \exists \Gamma_{\eta,t} \in \mathcal{K}, \inf_{p\geq 1} P(X_t^p \in \Gamma_{\eta,t}) \geq 1-\eta$$

- Condition (ii)

$$\exists C > 0, \forall T > 0, \forall t \in [0, T+1], \forall h \in [0, t],$$
$$E[|X^p_{t+h} - X^p_t|^2 | X^p_t - X^p_{t-h}|^2] \leq Ch^2$$

- Condition (iii)

$$\limsup_{\delta \to 0} \sup_{p \geq 1} E[|X^p_\delta - X^p_0|]^2 = 0.$$

To check these three conditions, the formula of the variance of X^p_t (equation A.2) is useful. Let us prove (i). $(X^p_\cdot)$ is a $\mathbb{R}$-value process, so (i) is equivalent to

$$\forall \eta > 0, \forall t \in \mathcal{Q}^*_+, \exists a_{\eta,t} \in \mathcal{R}^+, \inf_{p \geq 1} P(X^p_t > a_{\eta,t}) \leq \eta$$

Taking $a_{\eta,t} = \sqrt{\frac{t}{\eta}}$, the above property comes from Tchebychev's inequality.

We now prove (ii). Let $T > 0$ and $0 \leq t \leq T$

$$E[|X^p_{t+h} - X^p_t|^2 | X^p_t - X^p_{t-h}|^2] = E[|X^p_{t+h} - X^p_t|^2] E[|X^p_t - X^p_{t-h}|^2] = h^2$$

The second inequality comes from the independence of increments in the compound Poisson process.

Let us prove (iii). We have

$$\limsup_{\delta \to 0} \sup_{p \geq 1} E[\|X^p_\delta - X^p_0\|]^2 = \limsup_{\delta \to 0} \sup_{p \geq 1} \delta = \lim_{\delta \to 0} \delta = 0$$

■