THÈSE DE DOCTORAT DE MATHÉMATIQUES
DE L'UNIVERSITÉ DE GRENOBLE

préparée à l'Institut Fourier
Laboratoire de mathématiques

UMR CNRS 5582

Sur deux problèmes mathématiques de reconstruction phylogénétique

MIKAEL FALCONNET

Soutenue à Grenoble le 9 juillet 2010 devant le jury composé de :

Président	GÉRARD BESSON	Directeur de recherche, université de Grenoble
Rapporteurs	BERNARD PRUM	Professeur, université d'Évry
	MIKE STEEL	Professeur, university of Canterbury (NZ)
Examinateurs	AVNER BAR-HEN	Professeur, université Paris Descartes
	MANOLO GOUY	Directeur de recherche, université Lyon 1
	DIDIER PIAU	Professeur, université de Grenoble

Ce que l'on aime, on l'aime depuis toujours.

ANDRÉ HARDELLET

Remerciements

Tout d'abord, je remercie grandement Didier Piau pour avoir accepté d'être mon directeur de stage puis mon directeur de thèse. Pendant plus de trois ans, il m'a fait découvrir mon métier de jeune chercheur avec beaucoup de patience et de professionnalisme. Il m'a aiguillé et soutenu, il a été d'une disponibilité et d'une écoute extraordinaires, tout en sachant être rigoureux et exigeant avec moi comme avec lui-même. Grâce à lui, j'ai découvert les problématiques contemporaines de la phylogénie, et grâce à son expérience et à son recul, j'ai appris à nouer des dialogues avec des biologistes et à en extraire des problèmes mathématiques. Humainement, j'ai beaucoup apprécié la relation de maître à disciple que nous avons entretenue, le climat de confiance que nous avons maintenu et les discussions extra-mathématiques que nous avons pu avoir et qui ont renforcé le lien que nous avions. Je le remercie pour tout ça, et je sais que j'en oublie.

Je remercie sincèrement Bernard Prum d'avoir accepté la lourde tâche d'être un de mes rapporteurs, et je le remercie pour l'attention qu'il a portée à mon travail dans un délai plus que raisonnable.

It is quite an honor for me that Mike Steel accepted to be one of my referees, and I would like to express my most sincere thanks for his careful reading and helpful comments about my work.

C'est un plaisir et un honneur pour moi d'avoir dans mon jury Avner Bar-Hen et Manolo Gouy. C'est devant eux, entre autres, que j'ai fait mon premier exposé à la fin de mon M2, et je suis très fier de les retrouver pour cette soutenance de thèse, c'est une jolie manière de boucler la boucle.

Je trouve que Gérard Besson complète magnifiquement ce jury, et je suis très honoré qu'il ait accepté d'en faire partie. Tout d'abord parce qu'il s'est toujours enquis de l'avancée de mon travail, ce qui est assez rare pour être souligné, que sa bonne humeur m'a toujours fait du bien au sein du laboratoire.

Je tiens également à remercier les divers enseignants que j'ai eus quand j'étais étudiant à l'institut Fourier. J'ai une pensée particulière pour Agnès, Sandrine, Jean-Marc, Gérard, Jean, Christophe et Frédéric qui m'ont enseigné la théorie de la mesure et les probabilités, pour Stéphane qui m'a réappris à couper les epsilon en trois, et pour Michaël qui a été pour moi un enseignant exemplaire. J'ai également

une pensée pour Laurent qui a failli me faire pencher pour la géométrie différentielle. Je remercie également Hélène, Florent, Jean-Baptiste et Grégory qui m'ont drivé en prépa agrèg à Rennes et que j'ai toujours croisés avec plaisir au cours de cette thèse.

Maintenant, je remercie mes collègues doctorants pour les nombreuses discussions que nous avons eues au cours de ces trois années et pour les moments sympathiques que nous avons passés au laboratoire. Je remercie particulièrement Camille pour avoir relu une partie de mon premier chapitre, Bashar pour m'avoir aidé à préparer ma soutenance, et parce qu'ils le valent bien.

Je remercie également Laurence Després qui m'a aiguillé sur de bonnes références pour la rédaction de mon premier chapitre de thèse, et avec qui j'ai toujours plaisir à discuter.

Je remercie Brigitte, Géraldine, Martine, Gabrielle, Françoise, Ariane, Mickael et Hervé pour leur disponibilité et pour l'aide qu'ils m'ont fournie toute la durée de mon séjour à l'institut Fourier.

Enfin, je remercie les trois promotions d'étudiants qui ont accompagné agréablement cette thèse. C'est avec grand plaisir que j'ai été leur enseignant, et ils me laisseront un souvenir impérissable.

Je vais à présent remercier mes proches. Après avoir réfléchi, j'ai trouvé que le plus simple était de procéder de manière chronologique. En premier lieu, je tiens donc à remercier mes parents. C'est difficile de coucher sur le papier tout ce que je leur dois, mais je vais quand même essayer. Bien avant ma thèse, ils m'ont toujours soutenu, toujours encouragé, toujours aidé financièrement, et si j'ai pu me consacrer à mes études c'est grâce à eux. Dans les coups durs, c'est souvent vers eux que je me suis tourné, et c'était réconfortant de voir leur confiance aveugle envers ma réussite. Bien avant que je sois étudiant, ils m'ont inculqué des principes simples dont je ne saisis l'importance et l'impact que maintenant. Scolairement, je me suis débrouillé seul assez tôt, mais ils se sont occupés de moi quand j'étais petit, ils m'ont transmis le goût de la lecture, et ils ont toujours accordé de l'importance à ma réussite scolaire. Sans leur amour et leur dévouement, je n'aurais jamais fait cette thèse.

Mes parents avaient bien sûr d'autres occupations que de me faire réciter mes leçons, et ils m'ont donné une petite soeur, Jessika. Jusqu'à mon entrée au lycée, nous étions proches avec ma soeur. Après, j'ai été un peu plus absent et nous nous sommes moins vus. J'ai quitté ma soeur petite fille, et je l'ai retrouvée lycéenne quand je suis revenu à Grenoble pour commencer ma licence de maths. Cela m'a fait bizarre de passer des "Polly pocket" aux petits copains, mais je crois que c'est un peu de ma faute. Par contre, j'ai toujours senti son affection et son soutien, et je la remercie pour ça. Elle est toujours prête à m'accueillir dans son foyer, et j'ai toujours puisé du réconfort dans nos discussions. J'ai bien sûr une pensée pour tout le reste de ma famille, particulièrement pour mes oncle et tante Martine et Jean-Paul,

et pour ma grand-mère.

Jusqu'au CE2, j'étais scolarisé à Brest. Puis, mes parents ont déménagé pour retrouver leur village natal, et je suis allé à l'école de Montrigaud. J'ai rencontré dans cette école des amis qui sont toujours les miens, Yannis, Jean-Do (je leur réserve un paragraphe plus loin) et Mélanie. Mélanie était mon amoureuse en CM1, jusque là pas de quoi la glisser dans mes remerciements, mais elle a surtout été ma petite amie pendant près de six ans, et nous nous sommes quittés un peu avant qu'on ne m'attribue une allocation couplée. Elle m'a accompagnée pendant toutes mes études, puis nos chemins ont divergé avant le début de cette thèse. Sans elle, je ne suis pas sûr que je me serais tourné vers les mathématiques, et sans sa présence, je n'aurais jamais réussi mon concours d'entrée à l'ENS Cachan. Je la remercie pour tout ça. Je la remercie également d'avoir relu une partie de mon premier chapitre de thèse.

J'ai commencé à évoquer mes amis de Montrigaud, je vais tous les remercier à présent. Mélanie, Noelly, Renaud, Ludovic, Jean-Dominique, Jérémy et Yannis ont égayé tous mes week-end pendant de nombreuses années. Durant cette thèse, nos retrouvailles m'ont toujours procuré un immense plaisir, et j'y ai puisé à chaque fois une motivation supplémentaire. Ce sont tous des amis fidèles auxquels je tiens, leur compagne et compagnon aussi. Je tiens également à remercier Josette, Roger, Brigitte, Denis, Thérèse et Roland qui m'ont toujours accueilli chez eux à bras ouverts, et qui s'enquièrent toujours de mes nouvelles avec grand plaisir.

Après les amis de mon village, je tiens à remercier mes amis du lycée Albert Triboulet. Je pense en particulier à mes compagnons d'internat et à toutes les soirées qu'on a passées et qu'on continue de passer ensemble. Anthony, Sylvain, Sébastien, Éric, Yoann, Gwenn, Guillaume, Alexandre, Julien, Mey-Line et Gaëlle ont partagé leur quotidien avec moi pendant trois années, et nous avons forgé des liens qui sont toujours forts. Je tiens particulièrement à remercier Éric avec qui j'ai beaucoup passé de temps au O'Callaghan pendant ces trois années de thèse, et grâce à qui je ne perdrais sans doute jamais la brioche que j'ai accumulée lentement après toutes les pintes qu'on a ingurgitées. Au lycée, j'ai également fait la connaissance de Rémi, Édouard et surtout de Fabrice, qui m'a supporté pendant trois années de colocation. C'est un peu tôt pour éprouver ce qui va se passer lors de mon déménagement prochain, mais pour avoir quitté des compagnons d'internat par deux fois, je sais que ce sera sans doute douloureux. Je le remercie pour ces trois années qu'on a passées ensemble, pour le choix de nourriture saine et équilibrée qu'on a adopté, et pour les discussions profondes que nous avons eues sur la gent féminine. Je remercie également Rémi et Édouard de m'avoir accordé leur amitié.

Avec mon bac scientifique dans la poche, je suis finalement revenu sur les terres de mon enfance puisque j'ai intégré la classe préparatoire du lycée naval à Brest. Avant de remercier mes amis, je tiens à remercier les enseignants que j'ai beaucoup appréciés lors de mes trois années au KeuNeu. Christian Paisnel et Jean-Louis Guillerme ont été respectivement mes enseignants de mathématiques en sup et en

spé. C'est avec eux que j'ai découvert les mathématiques. Ils m'ont appris la rigueur, ils m'ont appris à raisonner, et ils ont su me faire aimer cette matière et en apprécier les beautés. Beaucoup de mes bagages, je leur les dois, et je les en remercie. Je remercie également Philippe Pillorget, mon prof d'anglais, et je crois que si je l'avais un peu plus écouté, je n'aurais pas tant galéré pour écrire cette thèse. Je tiens également à remercier Mademoiselle Andrieu, et Frédéric Bancel, mes enseignants de physique, qui m'ont appris à mener un calcul sans me poser de questions et à réfléchir après. Je sais, c'est sympa pour eux.

Je peux à présent évoquer mes amis du lycée naval. C'est parmi certains d'entre eux que j'ai pris la décision de me tourner vers les mathématiques et que j'ai renoncé à une carrière militaire. Ils m'ont accompagné, ils m'ont soutenu, pendant deux ans, trois pour certains. On a tout partagé, joies, peines, bitures, sanctions, concours, chambres, etc. Je me suis rendu compte récemment que ces amitiés étaient vraiment profondes, et que notre complicité est capable de revenir très vite, trop vite pour ceux qui sont à l'extérieur de nos souvenirs et qui se lassent de ces évocations qui n'ont aucun sens pour eux. En tous cas, je remercie VBB, Perkhut, Boyeau et de Cacq pour tous les souvenirs que nous partageons et je remercie en plus le Fruit, Carlo, Bridoux, OG-Hot et le Chwal pour m'avoir aidé à changer de voie et à affermir ma décision de venir en licence de mathématiques à Grenoble.

Quand j'ai intégré la prépa agrèg de Rennes, j'ai retrouvé un esprit de promo que j'aimais beaucoup. Durant cette année, j'ai particulièrement fait la connaissance de Benjamin, et depuis on a partagé beaucoup de moments ensemble, notamment deux écoles d'été de Saint-Flour. De temps en temps, ça nous arrive de parler de maths aussi et de discuter boulot. En tous cas, je le remercie pour tous les moments de détente qu'on a passés ensemble. Je remercie également Marc, Aurélien, Mathilde, Hélène, Kilian et Roland. Au cours de mon M2, j'ai apprécié l'amitié de Chloé, Anabel et Jérôme, et je regrette de ne pas avoir passé plus de temps avec eux.

C'est avec beaucoup de tendresse que je remercie Justine pour tous les beaux sourires qu'elle m'a donnés, et pour le temps que j'aurais dû consacrer à ma thèse et qu'elle m'a délicieusement fait perdre.

À présent, je voudrais remercier Antoine. J'ai eu énormément de plaisir à le côtoyer durant ces trois années, à échanger avec lui sur la thèse, sur la vie, et sur plein de choses en général. Sans lui, la thèse n'aurait pas été drôle du tout et je suis vraiment content d'avoir fait sa connaissance et de m'être lié d'amitié avec lui.

Enfin, last but not least, je voudrais remercier Olivier. Je me demande comment sera la recherche sans lui dans mon bureau. Après tout, j'ai passé énormément de temps en sa compagnie, que ce soit au bureau ou en dehors, et j'appréhende mon futur métier sans une amitié comme la sienne pour m'épauler. Ma thèse ne se serait sans doute jamais achevée sans son soutien et je le remercie pour tous les moments de complicité qu'on a passés ensemble. Les vrais savent.

Force et honneur.

Contents

Introduction

In this introduction, we present some motivations, our two main results, and the overall organization of the thesis.

Motivations

The reconstruction of the history of a given set of species is an important biological problem, since at least Darwin's inception of the theory of evolution. With the discovery of DNA as the genetic material, it became possible to base studies of evolution on molecular data. During the last decades, the development of statistical methods, the improvement of computer tools, and the huge accumulation of genetic data made these studies possible, and phylogenetics has known an important growth. Today, the applications of phylogenetics are numerous in several fields of molecular evolution including comparative genomics. For instance, phylogenetic trees can be used to predict the function of an unknown gene from its function in closely related species, see Eisen and Wu [EW02]. These applications require more and more accurate phylogenetic estimates and provide challenges at the junction of several areas: life sciences, stochastics, graph theory, combinatorics, computer science, and many others.

In this thesis we focus on one aspect of phylogeny, namely the methods of tree reconstruction in relation to the modeling of the evolution of DNA sequences.

Inferring distances for neighbour dependent substitution models

In most probabilistic models of the evolution of DNA sequences by nucleotidic substitutions, the evolution of each site is independent of the others and ruled by a Markovian kernel. With no interaction between different sites, the nucleotide at a given site converges in distribution to the stationary measure of the Markov chain whose dynamics is ruled by the 4×4 matrix of the substitution rates. Even more importantly perhaps, at equilibrium the sites become independent. As a conse-

quence, at equilibrium the frequency of every polynucleotide ought to be (at least roughly) equal to the product of the frequencies of its nucleotides.

In fact, biologists are well aware that the identity of the nucleotides in the immediate neighbourhood of a given site does affect, sometimes dramatically, the substitution rates at this site. For instance, in vertebrate genomes, the increased substitution rates of cytosine by thymine (CpG→TpG) and of guanine by adenine (CpG→CpA) in CpG dinucleotides are often quite noticeable and the chemical reasons of this CpG-methylation-deamination process are well known. A widely used quantification of the effect of this process is the so-called observed/expected ratio of CpG frequencies, denoted by CpGo/e, and defined as the ratio of the observed CpG frequency by the value of what the CpG frequency would be in the independent model, namely the product of the observed frequencies of C and G. As expected from the biochemical mechanisms involved, typical values of CpGo/e are < 1, the similarly defined ratios TpGo/e and CpAo/e are both > 1 and these depletion/excess effects are noticeable when the additional rates of CpG substitutions are high.

We now introduce the Jukes-Cantor model with CpG influence, hereafter denoted JC+CpG. This is the simplest non independent model of a class of models introduced and studied recently, denoted RN+YpR (we explain the notation later on), and designed specifically to take these effects into account.

JC+CpG is a continuous time model, such that the DNA sequences evolve under the combined effect of two mechanisms. The first mechanism is an independent evolution of the sites, like in the usual Jukes-Cantor model, where each substitution happens at the same rate, say at rate 1. A second mechanism is superimposed, which describes the substitutions due to the influence of the neighborhood. In JC+CpG, one assumes that the substitution rates of CpG→TpG and of CpG→CpA are both increased by a given additional rate, denoted by r, and that these are the only rates which are modified.

Hence, for instance, C→G at rate 1, C→A at rate 1 and C→T at rate 1 except when C belongs to the dinucleotide CpG, in which case C→T at rate $1+r$.

JC+CpG is the simplest model that we consider. We now introduce some notations needed to state our results.

The nucleotidic alphabet is $\mathscr{A} = \{A,C,G,T\}$ and the (bi-infinite) integer line is $\mathbb{Z}$. We consider two settings: either one observes two aligned sequences and one of these sequences is produced by the evolution mechanism described above, applied to the other sequence and running during a time t (we call this *the ancestral case*); or, one observes two contemporary aligned sequences, produced by two copies of the same ancestral unknown sequence evolving independently during a time t (we call this *the homologous case*).

In the ancestral case, one wants to estimate the elapsed time t. In the homologous case, one wants to estimate the divergence time t.

Let x and y be in $\mathscr{A}$. In the ancestral case, $(x,y)(t)$ denotes the frequency of sites occupied by x at time 0 (in the ancestral sequence) and by y at time t (in the present sequence). In the homologous case, $[x,y](t)$ denotes the frequency of sites occupied by x in one sequence and by y in the other one. Note that $[x,y](t) = [y,x](t)$ but that, a priori, there is no reason that $(x,y)(t)$ and $(y,x)(t)$ should coincide (indeed they do not, in general). Both $(x,y)(t)$ and $[x,y](t)$ are deterministic quantities which describe some theoretical frequencies and correspond to the comparison of infinite sequences.

For every x in $\mathscr{A}$ and $N \geqslant 1$, $(x,x)^N_{\text{obs}}$ and $[x,x]^N_{\text{obs}}$ denote the observed values of $(x,x)(t)$ and $[x,x](t)$ on aligned sequences of length N. In the ancestral case, $X(0) = (X_i(0))_{i\in\mathbb{Z}}$ denotes the ancestral sequence and $X(t) = (X_i(t))_{i\in\mathbb{Z}}$ denotes the present one. In the homologous case, the two sequences are $X(t) = (X_i(t))_{i\in\mathbb{Z}}$ and $X'(t) = (X'_i(t))_{i\in\mathbb{Z}}$. Then,

$$(x,x)^N_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{X_i(0) = X_i(t) = x\},$$

and

$$[x,x]^N_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{X_i(t) = X'_i(t) = x\}.$$

The estimators $(T^N_x)_{\text{obs}}$ of the elapsed time t and $[T^N_x]_{\text{obs}}$ of the divergence time t, based on the observation of aligned sequences of length N, are defined as the solutions in t of the equations

$$(x,x)(t) = (x,x)^N_{\text{obs}} \quad \text{and} \quad [x,x](t) = [x,x]^N_{\text{obs}}, \quad \text{respectively.}$$

We prove the following result.

Theorem. *In JC+CpG, assume that the ancestral sequence is at stationarity. Then, for every x in $\mathscr{A}$, there exist explicit observed quantities $(\alpha^N_x)_{\text{obs}}$ and $[\alpha^N_x]_{\text{obs}}$, such that*

$$(\alpha^N_x)_{\text{obs}}\sqrt{N}((T^N_x)_{\text{obs}} - t) \quad \textit{and} \quad [\alpha^N_x]_{\text{obs}}\sqrt{N}([T^N_x]_{\text{obs}} - t)$$

both converge in distribution to the standard normal law when $N \to +\infty$.

Formulas for $(\alpha^N_x)_{\text{obs}}$ and $[\alpha^N_x]_{\text{obs}}$ are in section 3.2.

This theorem yields asymptotic confidence intervals for the elapsed time between an ancestral sequence and a present one, and for the time of divergence between two present sequences issued from a common ancestral one, based on the observations $(x,x)^N_{\text{obs}}$ and $[x,x]^N_{\text{obs}}$, respectively.

The strategy to prove this theorem is the following. Both estimators $(T^N_x)_{\text{obs}}$ and $[T^N_x]_{\text{obs}}$ are based on the theoretical proportion of identical nucleotides of type x in the two sequences, denoted by $(x,x)(t)$ and $[x,x](t)$ respectively. We show that

these are decreasing functions of time t. The proof for $x = C$ (the case $x = G$ being similar) relies on two ideas. First, we note that the choice of a special (reduced) alphabet to encode dinucleotides provides an autonomous Markovian evolution of these encoded dinucleotides, with a 4×4 rate matrix. As a consequence, we are able to compute an explicit expression for $(C,C)(t)$ and to prove that $t \mapsto (C,C)(t)$ is a decreasing diffeomorphism. Second, a reversibility argument provides the equation $[C,C](t) = (C,C)(2t)$, which implies trivially that $t \mapsto [C,C](t)$ is a decreasing diffeomorphism. Note that for $x = A$ (the case $x = T$ being similar), the relation $[A,A](t) = (A,A)(2t)$ is false. For $x = A$, we prove that $t \mapsto (A,A)(t)$ is a decreasing diffeomorphism but we have to leave open the case of $t \mapsto [A,A](t)$. However, subsection 5.1.1 contains a possible route to prove that $t \mapsto [A,A](t)$ is indeed a decreasing diffeomorphism, as our simulations suggest.

One sees that the study of the random variables $(C,C)^N_{\text{obs}}$ and $[C,C]^N_{\text{obs}}$ provides results for $(T^N_C)_{\text{obs}}$ and $[T^N_C]_{\text{obs}}$ through an inversion of functions, the delta method, and Slutky's lemma. Relying on the special dependency structure of the JC+CpG model, we compute explicitly the mean and the variance of $(x,x)^N_{\text{obs}}$ and $[x,x]^N_{\text{obs}}$ for every x, and we provide central limit theorems (with explicit variances) for these quantities.

In the general case of an RN model with YpR influence, we extend the result above under a proviso, namely that the equation defining the estimator has a unique solution. This proviso requires to prove that $t \mapsto (x,x)(t)$ and $t \mapsto [x,x](t)$ are decreasing diffeomorphisms, and this is still open, even if some simulations support this conjecture.

An unfortunate aspect of Bayesian methods in phylogeny: the star paradox

Bayesian inference in phylogeny is a powerful tool to infer trees. However, one must be careful about some unfortunate mathematical aspects of the method. We study one of these, called the Bayesian star paradox.

This paradox refers to the fact that a given resolved tree can be highly supported even when the data is generated by an unresolved star tree. Recent studies highlight the fact that the paradox can occur in the simplest setting, namely, for an unresolved rooted tree on three taxa and two states, see Yang and Rannala [YR05] and Lewis and al. [HLH05]. Kolaczkowski and Thornton [KT06] presented some simulations and suggested that artifactual high posteriors for a particular resolved tree might disappear for very long sequences. Previous simulations in [YR05] were plagued by numerical problems, which left unknown the nature of the limiting distribution on posterior probabilities.

The statistical question underlying the star paradox is to determine whether or not the Bayesian posterior distribution of the resolutions of a star tree converges to

the uniform distribution, almost surely, when the length of the sequence tends to infinity. In a recent paper, Steel and Matsen [SM07] disprove this almost sure convergence in the simplest non trivial case, namely for three taxa and a binary alphabet, thus ruining Kolaczkowski and Thornton's hope. Steel and Matsen's result holds for a specific class of branch length priors, which they call *tame*. We now describe their setting.

One considers binary sequences of length n generated by a star tree R_0 on three taxa with strictly positive edge length t. Let $N_{0:3}$ denote the resulting data, summarized by four site pattern countings summing to n. Consider the three resolved trees R_1, R_2 and R_3 drawn below.

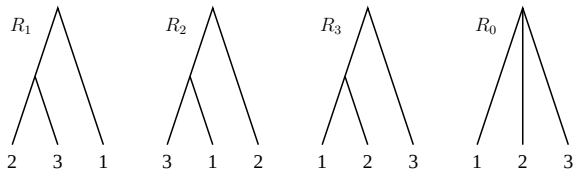


Figure 1: For every i in $\{1,2,3\}$, i is the outlier in tree R_i. The star tree is R_0.

Every tree R_i with i in $\{1,2,3\}$ is entirely described by a topology and by two branch lengths T_{i} and T_{e}. The *internal* length T_{i} describes the time elapsed between the two speciations and the *external* length T_{e} describes the time elapsed since the last speciation, hence the common ancestor is at distance $T_{\mathrm{e}} + T_{\mathrm{i}}$ in the past.

Say that a distribution of (T_e, T_i) is *tame* if it has a smooth joint probability density function that is bounded and everywhere non zero. For instance, if T_{e} and T_{i} are independent and exponentially distributed, their distribution is tame. Steel and Matsen proved the following result.

Theorem (Steel and Matsen [SM07])**.** *Consider any prior on the three resolved trees R_1, R_2 and R_3 and any tame prior distribution on their branch lengths. Then, for every positive ε, there exists a positive δ such that, when n is large enough, for every $i \in \{1,2,3\}$,*

$$\mathbb{P}\left(\mathbb{P}(R_i|N_{0:3}) \geqslant 1-\varepsilon\right) \geqslant \delta.$$

Later on, Steel and Matsen's theorem was taken into account by Yang [Yan07] and reinforced by theoretical results on the posterior probabilities by Susko [Sus08].

Our main result is that Steel and Matsen's conclusion holds for a wider class of priors for branch lengths, which we call *tempered*. The definition of the class of *tempered* distributions is in section 4.2. Since it is rather involved, we provide some concrete examples after the statement of the result.

Theorem. *The conclusion of Steel and Matsen's theorem above holds for every tempered prior distributions on branch lengths.*

Every tame prior is tempered but the converse is false. First, the condition to be tempered involves the cumulative distribution function of (T_e, T_i) and not the density of their distribution, which may not exist. For instance, tempered prior distributions may incorporate accumulations of Dirac masses. Furthermore, continuous prior distributions exist, which are tempered but not tame.

For instance, assume that T_e and T_i are independent, that T_e is exponentially distributed and that the distribution of T_i is either uniform on an interval $[0, \vartheta]$, with $\vartheta > 0$, or a power distribution $\kappa t_i^{\kappa-1} dt_i$ of the interval $[0,1]$, with κ in $]0,1]$. Then the distribution of (T_e, T_i) is tempered but not tame.

Plan of the thesis

We wrote chapter 1 as an introduction to molecular genetics and phylogenetics, aimed at readers not so familiar with some basics of biology. We do not pretend to give an exhaustive review of genetics, only a personal view of the notions which are necessary to understand the rest of the thesis. We also introduce some vocabulary related to phylogenetic trees, and detail some distance-based reconstruction methods.

In chapter 2, we describe some nucleotidic substitution processes. We detail two independent models: the Jukes-Cantor model and the Kimura model, and we recall how to provide consistent estimators of genetic distances for DNA sequences evolving under these models. We also provide a description of a class of neighbour dependent substitution processes introduced in [BGP08] and called RN+YpR.

In chapter 3, we prove that one can compute consistent estimators for DNA sequences evolving under the Jukes-Cantor model with CpG influence, the simplest non trivial model in RN+YpR. We also show how to extend these results to the whole class, assuming that some technical properties hold. Recall that this is the first step needed to build phylogenetic trees based on any RN+YpR model.

We mention that the content of this chapter is the subject of *Phylogenetic distances for neighbour dependent substitution processes* [Fal10], a paper now published in *Mathematical Biosciences*.

In chapter 4, we introduce briefly the use of Bayesian inference in phylogeny, we describe one of its unfortunate aspects named the star paradox, and we prove that the result due to Steel and Matsen and recalled above occurs for a wide class of prior distributions.

We mention that the content of this chapter is the subject of *Priors for the Bayesian paradox* [Fal09], a paper now in revision for *Mathematical Biosciences*.

Finally, we briefly present in chapter 5 some further lines of research, including natural follow ups of the results of the thesis, as well as a somewhat more ambitious project aiming at developing some new models of evolution.

Chapter 1

A short introduction to molecular genetics and phylogenetics

To understand the origin of the mathematical questions treated in this thesis, some background in genetics is needed. We present this in the first section, aiming at readers who are not so familiar with biology. Second, we introduce the reader to phylogenetics and to phylogeny reconstruction with distance-based methods.

Some explanations are in order here. We chose to present some classical distance-based methods for phylogeny reconstruction for the following reasons. First, since one theme of this thesis is the inference of distances for homologous DNA sequences, our context is the construction of phylogenetic trees from distances between DNA sequences. Second, even if these methods are well documented, and if the mathematical ideas behind them are not so complicated, we personally experienced some difficulties when faced with one of these algorithms. For instance, it was not clear to us how and when the algorithm should end. With the idea that some readers might encounter similar difficulties, we chose to provide complete examples of two classical distance-based methods. In other words, an in-depth understanding of distance-based methods is not required to understand the results of this thesis, and we do not pretend to provide a complete nor thorough review of the subject, but, since inferring distances is only one step in the process of phylogeny reconstruction, we felt important to provide a grasp of the rest of the process as well.

For this introduction to biological concepts and results, we used, among other sources, [GL00], [SPAP95a], [SPAP95b], [Yan06], [Gas05] and [GS07].

1.1 About molecular evolution

In this section, we recall some aspects of the historical discovery of the genetic material, mainly the deoxyribonucleic acid (DNA), of the birth of molecular evolution related to the chemical structure of DNA, and of the possible modifications of DNA sequences.

1.1.1 Some historical moments of molecular evolution

The concept of evolution for living organisms is now widely accepted among scientists and laypersons, even if some skeptical communities remain. The situation was almost exactly the opposite before the advent of Darwin's theory and I like to compare the fight for evolution to Copernic's fight against geocentrism. I think that some dates and some names deserve to be mentioned in any thesis about phylogenetics.

A brief history of the concept of evolution before Darwin

The modern concept of evolution was introduced by Darwin[1] (photograph below) in the middle of the nineteenth century in his famous book *The Origin of Species*. His theory was that populations evolve over the course of generations through a process of natural selection, and that all species of life descended over time from common ancestors. We now briefly discuss the positions of some of his predecessors.

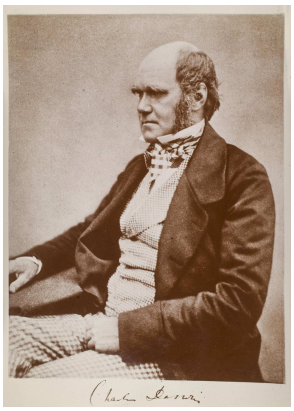


Figure 1.1: Charles Darwin

The origin of animals and men has always been a nagging question for mankind. The Creation-Evolution controversy began to be particularly acute during the eigh-

[1] Charles Robert Darwin (February 12, 1809 – April 19, 1882).

teenth century. Indeed, the existence of fossils showing past extinctions was an uncomfortable discovery for some widely held beliefs, for instance in the context of Judeo-Christian religions. The influence of the ideas of Enlightenment could also explain the birth of different thinkings. One must be careful when one talks about this period, because it appears difficult to understand the influence and thinking of people. For instance, Cuvier[2], one of the fathers of comparative anatomy, supported the theory of catastrophism[3] to explain these extinctions, a theory confortable for the Church, but he never talked about religion. A few decades before him, Buffon[4] thought differently than him, but he stayed careful to not seem too much embarrassing.

However, at the beginning of the nineteenth century, a theory of gradual changes in living organisms was advocated by Lamarck[5], who introduced the first coherent theory of evolution and refused the theory of catastrophism. One can wonder why Lamarck is less famous than Darwin. First, Lamarck's theory was only based on observations and thinking, and not on experimentation. As a consequence it was much harder to defend. Second, his theory was (quite) wrong. Lamarck thought that living organisms were able to adapt to the pressures of their environment during their lifetime, and then to transmit these acquired abilities to their offsprings. Until recent discoveries in the field of molecular evolution, biologists thought that this theory, named Lamarckism after him, was entirely wrong. Now, we know that there exists some cases where organisms adopt such a process, see [NLS+09], but that these are rare.

Lamarck's work was praised by Darwin, because it aroused interest about possible scientific explanations for changes in organism. But the real father of evolutionism remains Darwin, even if some scientists now think that the influence of natural selection is less important than Darwin thought.

DNA is the genetic material

One sees that evolutionism began with Darwin. He convinced people that living organisms evolve over the course of generations, but he did not know precisely how the transfer of biological information occurs.

A few years after the publication of *The Origin of Species*, Mendel[6] established

[2]Georges Léopold Chrétien Frédéric Dagobert Cuvier (August 23, 1769 – May 13, 1832), French naturalist and zoologist.

[3]Catastrophism is the idea that Earth has been affected in the past by sudden, short-lived, violent events, possibly worldwide in scope.

[4]Georges-Louis Leclerc, Comte de Buffon (September 7, 1707 – April 16, 1788), French naturalist and mathematician. Buffon's needle problem is the earliest geometric probability problem to be solved.

[5]Jean-Baptiste Pierre Antoine de Monet, Chevalier de la Marck (August 1, 1744 – December 18, 1829), French soldier, naturalist, academic.

[6]Gregor Johann Mendel (July 20, 1822 – January 6, 1884), Augustinian priest and scientist.

Figure 1.2: Jean-Baptiste Lamarck

in *Experiments in Plant Hybridization* the existence of elementary "characters" of heredity, and the statistical laws governing their transmission from one generation to the next. However, he knew nothing about the nature of these "characters", and was not sure of the existence of a physical material which would contain this information. In 1869, Miescher[7] discovered a phosphorus-containing substance into the nuclei of white blood cells, which he called "nuclein" and which was, later on, renamed as "nucleic acid". Miescher thought for a while that this substance might be related to heredity, but finally changed his mind. After Miescher, several biologists increased the knowledge on deoxyribonucleic acid, its structure and its possible role in heredity. Until 1944 and the Avery-MacLeod-McCarty experiment[8], most scientists thought that the genetic information was carried by proteins. This hypothesis was abandoned in 1952 with the Hershey-Chase experiments[9].

The birth of molecular evolution

The field of molecular evolution was at this moment wide opened, and three fundamental directions appeared.

- First, the classification of the living world and the reconstruction of the evolutionary history could be based on molecular data, and not anymore on traditional fields only. Molecular phylogenetics was born.

Mendel gained posthumous fame as the figurehead of the new science of genetics for his study of the inheritance of certain traits in pea plants.

[7]Johannes Friedrich Miescher (August 13, 1844 - August 26, 1895), Swiss biologist.

[8]The Avery–MacLeod–McCarty experiment was an experimental demonstration, reported in 1944 by Oswald Avery, Colin MacLeod, and Maclyn McCarty, that DNA is the substance that causes bacterial transformation.

[9]The Hershey–Chase experiments were a series of experiments conducted in 1952 by Alfred Hershey and Martha Chase, confirming that DNA was the genetic material.

- Second, since one knew where the genetic information was located, it became possible to study the mechanisms of the changes in the genetic material.

- Last but not least, the question of the origin of life could be studied with a new point of view.

Of course, these questions are related, and a step in the comprehension of changes is a step for phylogenetics. We expose now some information related to the genome and its role in the transmission of heredity.

1.1.2 Some basic knowledge about the genome

The hereditary information of almost all living organisms, some viruses excepted, is carried by deoxyribonucleic acid (DNA) molecules. We discuss briefly the chemical structure of DNA, and the process of DNA replication.

Chemical structure

Chemically, DNA consists of two complementary strands twisted around each other to form a right handed double helix, illustrated in figure 1.4. This structure was suggested by Watson[10] and Crick[11], in 1953, who based their molecular model on a single X-ray diffraction[12] image taken by Franklin[13] and Gosling[14] in 1952.

Each strand of DNA is a long linear succession of repeating units, called nucleotides, which are of four possible kinds. Two of these kinds are purines, adenine (A) and guanine (G), and two are pyrimidines, thymine (T) and cytosine (C).

[10]James Dewey Watson (born April 6, 1928), American molecular biologist. He, Francis Crick, and Maurice Wilkins were awarded the 1962 Nobel Prize in Physiology or Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material".

[11]Francis Harry Compton Crick (June 8, 1916 – July 28, 2004), British molecular biologist, physicist, and neuroscientist. He, James D. Watson and Maurice Wilkins were jointly awarded the 1962 Nobel Prize for Physiology or Medicine "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material".

[12]X-ray scattering techniques are a family of non-destructive analytical techniques which reveal information about the crystallographic structure, chemical composition, and physical properties of materials and thin films. These techniques are based on observing the scattered intensity of an X-ray beam hitting a sample as a function of incident and scattered angle, polarization, and wavelength or energy.

[13]Rosalind Elsie Franklin (July 25, 1920 – April 16, 1958), British biophysicist, physicist, chemist, biologist and X-ray crystallographer. She is still best known for her work on the X-ray diffraction images of DNA. Her data, according to Francis Crick, was "the data we actually used" to formulate Crick and Watson's 1953 hypothesis regarding the structure of DNA.

[14]Raymond Gosling (born 1926), British biophysicist. He was a research student of Rosalind Franklin.

Figure 1.3: James Watson, Rosalind Franklin, and Francis Crick

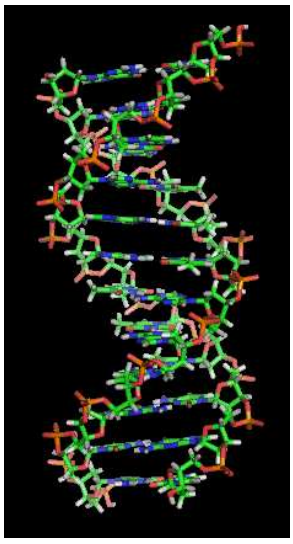


Figure 1.4: A section of DNA

To give an idea of the length of a DNA molecule in different organisms, and the information that it represents, we indicate some genome sizes (see [Ped71], [Fa76], [Ba97], [Ga96] [Ta06], [Aa00], [La01]) in table 1.5, and corresponding sizes in terms of books. Note that techniques for measuring the genome size existed at the beginning of the 1950s whereas the first DNA sequencing[15] occurred in 1976 only, and concerned the bacteriophage MS2 RNA whose genome length is 3659 bp, see [Fa76]. The first sequencing of the human genome was achieved in 2001, and cost three billion dollars (one dollar per nucleotide). In 2007, Watson's genome was sequenced with a new machine and some new techniques, at a total cost of less than one million dollars.

The backbone of one DNA strand is made of alternating phosphate and sugar (deoxyribose) residues. The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar

[15] DNA sequencing is the process of determining the nucleotide order of a given DNA fragment.

Organism type	Organism	Genome size	Correspondence
Virus	Bacteriophage MS2	3.5 10^3 bp	0.15 page
Bacterium	Escherichia coli	4.6 10^6 bp	168 pages
Yeast	Saccharomyces cerevisiae	12.1 10^6 bp	484 pages
Plant	Populus trichocarpa	480 10^6 bp	12 volumes
Insect	Drosophila melanogaster	130 10^6 bp	4 volumes
Mammal	Homo sapiens	3 10^9 bp	80 volumes
Fish	Protopterus aethiopicus	130 10^9 bp	3440 volumes

Figure 1.5: Writing the nucleotide sequence of genetic material with 25 kb per page and 1500 pages per volume.

rings. These asymmetric bonds mean a strand of DNA has a direction. By convention, a DNA sequence is written in the order of transcription, from the $5'$ to the $3'$ end as indicated in figure 1.6. As a consequence the complementary strand of the leading strand, called the lagging strand, is oriented in the opposite direction.

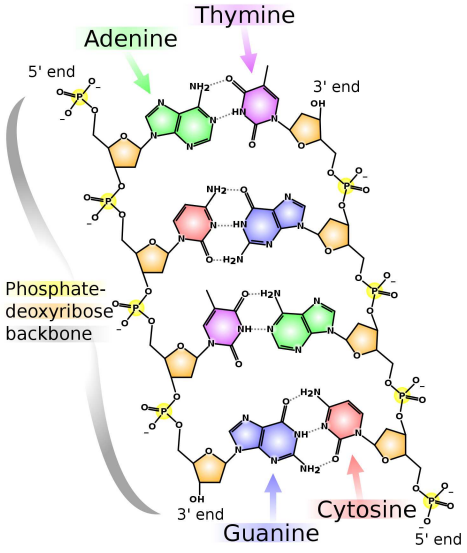


Figure 1.6: Chemical structure of DNA

DNA replication and heredity

The process of DNA replication is paramount to all life as we know it. This process occurs before each and every cell division and consists in the copy of one double-stranded DNA molecule into two identical DNA molecules. Its principle is illustrated in figure 1.7.

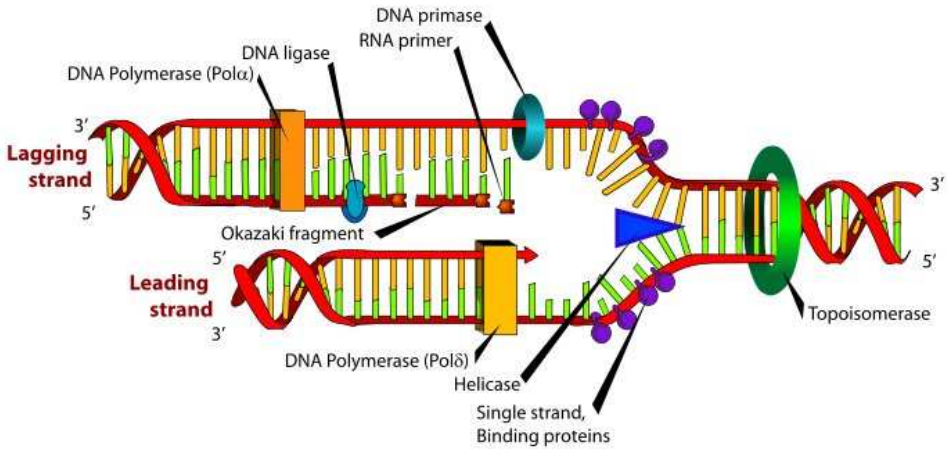


Figure 1.7: DNA replication is the process of copying a double-stranded DNA molecule

Thus, the offspring cells keep the same characteristics than their parent cell, and, through generations, the instructions contained in the DNA such as protein-coding genes[16] are passed from parent cells to offspring cells.

1.1.3 Mutations enter the stage

DNA replication is not perfect, in the sense that some replication errors occur during the process, and as a consequence new DNA sequences appear. If the error occurs in a somatic cell[17], the mutation will not be inherited. However, if the error happens in a germline cell[18], the mutation is transmitted to the offspring, and this is more interesting from the evolutionary point of view.

Errors during the DNA replication are not the only cause of mutations. Changes can also be caused by radiations, viruses, recombinations, and many other causes.

DNA sequences can be altered in a number of ways during their replications. Here are some of them.

Substitutions exchange a single nucleotide for another. These are classified as transitions on the one hand, which exchange a purine for a purine ($A \leftrightarrow G$) or a pyrimidine for a pyrimidine ($C \leftrightarrow T$), and transversions on the other hand, which

[16] A gene used to be defined as a segment of DNA that codes for a polypeptide chain or specifies a functional RNA molecule, but recent studies make the definition more complex.

[17] Somatic cells (diploid) are the cells which form the body of an organism, as opposed to germline cells.

[18] The germline of a mature or developing individual is the line (sequence) of germ cells that contain genetic material that may be passed to a child. For example, sex cells such as the sperm or the egg, are part of the germline.

exchange a purine for a pyrimidine or a pyrimidine for a purine ($C/T \leftrightarrow A/G$).

Insertions add one or more extra nucleotides into the DNA. On the contrary, **deletions** remove one or more nucleotides from the DNA. Note that insertions can be reverted by excision of the transposable element, whereas deletions are generally irreversible.

Recombination is a process by which a molecule of DNA is broken and then joined to a different DNA molecule. It can occur in meiosis[19] as a way of facilitating chromosomal crossover[20].

Islands of mutations and the dinucleotide CpG

Modeling or even describing mutations is a difficult task. Indeed, these depend on too many parameters: the nature of the living organism, the nature of the DNA (mitochondrial[21] or not), the location in the DNA sequence (coding or non-coding regions), etc. As a consequence, we cannot even pretend to explain all these aspects here. Rather, we insist on a particular point which motivates this thesis.

One knows that regions of the DNA are more prone to mutations than others and that some of them are related to the dinucleotide CpG. Here and later on in this thesis, we use the notation "CpG" as a shorthand for "$5' - CG - 3'$". In CpG dinucleotides of mammalian genomes, the cytosine is frequently methylated[22] (see [Bir80]). Methylated CpG dinucleotides may change into TpG with higher frequency, and consequently into CpA on the complementary strand. There exist other hotspots in DNA, but this one is of particular interest to the biologist because of the existence of CpG islands in mammalian genomes (see [Bir86], [AB91b], [AB91a]). A CpG island is a region of DNA that have a higher concentration of CpG sites, and frequently these regions are functional ones (see [AB99]). This suggests that the process of methylation is repressed in such regions, and the comprehension of this phenomenon is important in biology. In section 2.3, we will see how the phenomenon can be taken into account in mathematical models.

We do not go deeper about this brief presentation of DNA. To understand properly the importance of DNA in evolution studies, the reader should keep in mind that DNA has the role of a very important mechanism of storage of information and that it is used constantly by living organisms, and systematically copied to transmit information to the next generation. As a consequence, the DNA of every living

[19] Meiosis is a process of reductional division in which the number of chromosomes per cell is divided by two. Meiosis is essential for sexual reproduction and therefore occurs in all eukaryotes that reproduce sexually.

[20] Crossing over is an exchange of genetic material between homologous chromosomes.

[21] Mitochondrial DNA (mtDNA) is the DNA located in some organelles called mitochondria. These are structures within cells that convert the energy from food into a form that cells can use. Most other DNA present in eukaryotic organisms is found in the cell nucleus.

[22] Cytosines in CpG dinucleotides are methylated by DNA methyltransferases in many eukaryotic organisms to form 5-methylcytosine. In mammals, 70% to 80% of CpG cytosines are methylated.

organism can be seen as a historical record of distinctive marks of evolutionary processes, and one can (in theory) deduce the chronology of evolution from this molecular data.

1.2 Molecular phylogenetics

Phylogenetics may be defined as the study of evolutionary relationships among living organisms. Traditional approaches to the study of the historical records of evolution include morphology[23], anatomy[24], physiology[25], and paleontology[26].

Even if none of these approaches is abandoned, molecular data now available is a most suitable basis for this study, and this area of phylogeny is called molecular phylogenetics. Both traditional and molecular phylogenetics are supported by the same idea: closely related organisms in evolution have a high degree of agreement in physical characters or in their molecular structure.

In this section, we discuss some goals of phylogeny, and we detail two distance based algorithms used to reconstruct phylogenetic trees.

1.2.1 Descriptive tools

One purpose of phylogenetics is to trace the common ancestry of organisms living today, typically by reconstructing phylogenetic trees. We discuss some basic concepts used to describe these objects.

Bifurcating trees

Imagine that one wishes to trace the history of three species. Going back in time, one can hope to link two of these species sharing a common ancestor; going further back in time, one can hope to find a common ancestor of the three species. The construction of these successions of branching for species is the purpose of phylogeny, and motivates the representations below.

Definition 1.2.1. *A* graph *is an ordered pair* (V,E) *where* V *is a set of objects called* vertices *or* nodes*, and* E *is a set of objects called* edges *or* branches*, each connecting two vertices. A* path $(v_0, v_1, \ldots, v_k)$ *is a sequence of elements of* V *such that for every integer* $0 \leqslant i < k$*,* (v_i, v_{i+1}) *is an edge. A* cycle*, also called a* loop*, is*

[23] Morphology is the study of the form, structure and configuration of organisms.

[24] Anatomy is a branch of biology and medicine that is the consideration of the structure of living things.

[25] Physiology is the science of the functioning of living systems.

[26] Paleontology is the study of prehistoric life, including organisms evolution and interactions with each other and their environments.

a path $(v_0, v_1, \ldots, v_k)$ *such that* $k > 2$, $v_0 = v_k$ *and* $v_i \neq v_j$ *for every* $0 \leqslant i, j < k$. *A* tree *is a connected graph without loop.*

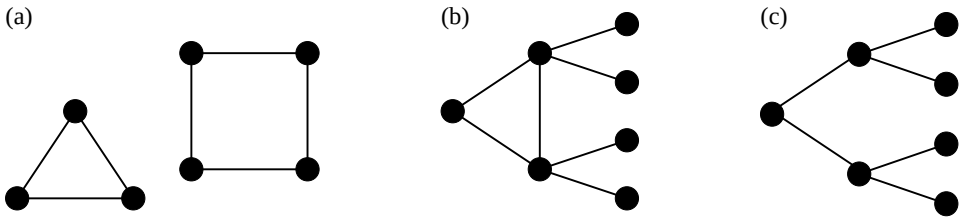


Figure 1.8: (a) Unconnected graph. (b) Connected graph with loop. (c) Tree.

Phylogenetic trees should show the historical evolution of species. Hence their representation is codified to make easier the reading of time and diversity, as follows.

Definition 1.2.2. *The* leaves *(external nodes) represent present-day species, often named taxa*[27]. *The internal nodes represent extinct ancestors for which no sequence data are available. The ancestor of all taxa is the* root *of the tree.*

This is illustrated in figure 1.9. A rooted tree has an orientation from the past (the root) to the present (the leaves), and the unique path between the root and a leaf represents the different speciations of one present species.

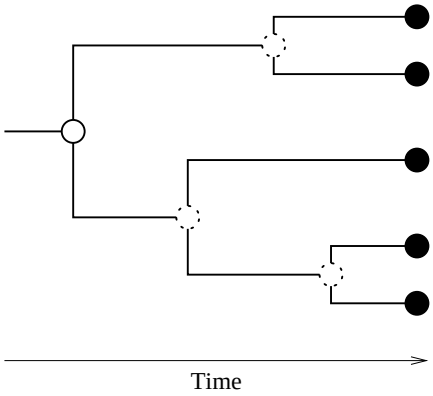


Figure 1.9: A phylogenetic tree. Filled, dashed and solid circles denote respectively external nodes, internal nodes and the root.

Note that one can also use unrooted trees. This erases the direction of the arrow of time but the tree still represents the diversity and closeness between species.

[27] A taxon (plural: taxa) is a group of (one or more) organisms, which a taxonomist adjudges to be a unit.

In figure 1.9 every internal node is incident to exactly three branches, two derived and one ancestral. Indeed, in evolutionary studies, one commonly assumes that the process of speciation is binary. Thus, the common representation of phylogenetic trees uses bifurcating trees, in which each ancestral taxon splits into two offspring taxa. However, when the speciation events are uncertain, because of a lack of data or of inacurracies of the method, one may have to use trees with three or more offspring taxa for a given ancestral taxon.

We now discuss the main characteristics of phylogenetic trees.

Branching patterns and branch lengths

The branching pattern of a tree is called the topology of the tree. For example, for three species, there are three possible bifurcating trees, and four if one adds the star tree, as illustrated in figure 1.10. In terms of evolution, the branching pattern in figure 1.10 (a) means that a speciation occurred between the species C and the common ancestor of species A and B, and that another speciation occurred later, which yielded species A and B.

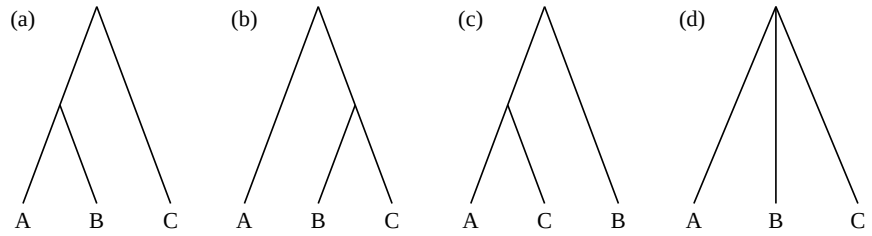


Figure 1.10: Possible rooted trees for three species A, B and C.

The number of possible trees increases (more than) exponentially with the number of species. To understand why, see figure 1.11 which illustrates the stepwise addition algorithm for unrooted trees introduced in [CSE67]. The algorithm is the following.

One starts with the single unrooted tree for 3 species. The fourth species can be added to each of the 3 branches of the tree. Thus, there are 3 unrooted trees for 4 species. Each tree on 4 species has 5 branches to which the fifth species can be added. Thus, there are 5×3 trees on 5 species. Likewise, a tree on $n-1$ species has $2n-5$ branches (since adding a species to a tree means adding 2 branches), to which the nth species can be added. Hence there are $u_n = (2n-5) \times (2n-7) \times \cdots \times 5 \times 3$ unrooted trees for n species.

To work out the number of rooted trees for n species, note that each unrooted tree has $2n-3$ branches, and that the root can be placed on any of these branches. This generates $2n-3$ rooted trees for every unrooted tree, thus the number of rooted trees for n species is the number u_{n+1} of unrooted trees for $n+1$ species.

Numerical values are $u_{11} \approx 35$ millions rooted trees on 10 species and $u_{14} \approx 14$ billions rooted trees on 13 species.

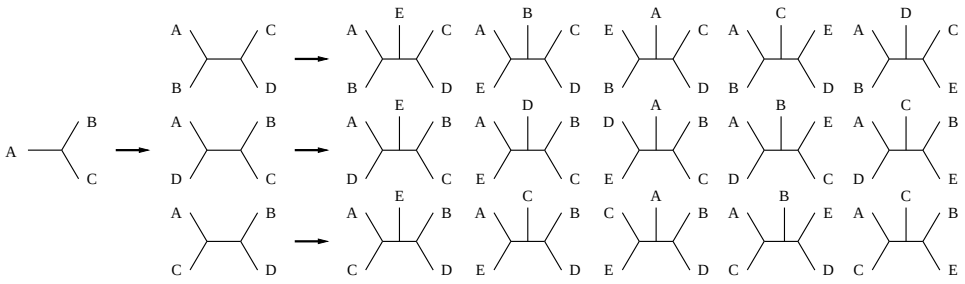


Figure 1.11: Illustration of the stepwise addition algorithm.

Definition 1.2.3. *A* cladogram *is a tree topology. A* phylogram *is a tree topology with branch lengths. A* dendrogram *is a rooted phylogram where the length of the path from the root to every leaf is the same.*

One uses dendrograms when all the species evolve at the same speed, then branch lengths represent time durations. Otherwise one uses phylograms, and then branch lengths represent diversities between species. In chapter 2, we define distances between species through their DNA sequences, and use these distances to quantify differences between species.

To underscore the importance of branch lengths, we refer to figure 1.12, showing two possible dendrograms for species A, B and C related by the same cladogram. In case (a), the phylogeny is highly resolved hence one is pretty sure that A and B share a common ancestor which is not an ancestor of C. In case (b), the tree is close to the star tree drawn in figure 1.10 (d), hence the order of the speciation events between species A, B and C is uncertain.

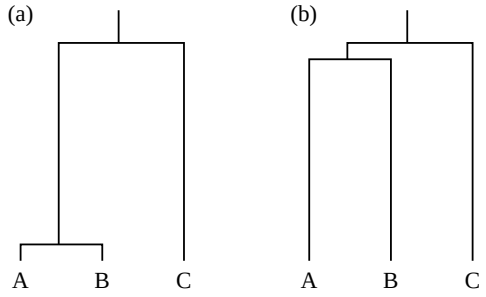


Figure 1.12: One branching pattern: two possible phylograms for Human, Chimpanzee and Macaque.

Assume that a given collection of present day species share a common ancestor and that the evolution since this common ancestor is ruled by a *true* tree. One objective

of phylogenetics is to reconstruct this tree, its branching patterns and the branch lengths. We now present some methods to infer this tree from molecular data.

1.2.2 Distance-based methods

In distance-based reconstruction methods, one computes distances from pairwise comparisons between sequences, based on nucleotidic substitution models. Hence, one considers that the only mutations that can occur are substitutions. A presentation of these models and of the computation of distances are in chapters 2 and 3. To summarize this step, starting from n species one computes $\frac{1}{2}n(n-1)$ values representing the distances between each pair of species.

Before presenting some distance based methods, we explain some difficulties associated to the relation between trees and distances.

Trees versus distances

Consider as an example the dendogram T_{ex} on five species A, B, C, D and E represented in figure 1.13 below.

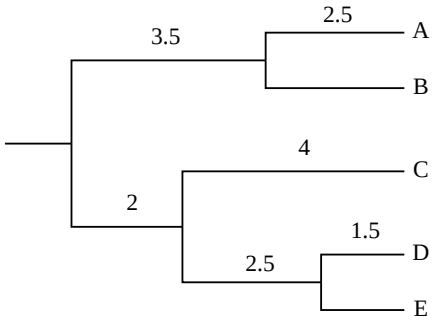


Figure 1.13: Dendrogram T_{ex}

The distance between two species is the sum of the lengths of the edges on the path between them, for instance the distance between C and D is $4+2.5+1.5=8$. We summarize these distances in the matrix Δ_{ex} below.

$$\Delta_{ex} = \begin{array}{c|cccc} & A & B & C & D \\ \hline B & 5 & & & \\ C & 12 & 12 & & \\ D & 12 & 12 & 8 & \\ E & 12 & 12 & 8 & 3 \end{array}$$

Distances between species are easy to compute from the dendrogram. We are interested in the reverse process, which is to construct a phylogram, and possibly a dendrogram, corresponding to a given set of distances between species.

First, to be able to draw a dendrogram assumes a condition on the set of distances, called the *three-points condition*.

Proposition 1.2.4 (Three-points condition)**.** *Consider a dendrogram T and the corresponding distance d. Then, for every leaves x, y, and z,*

$$d(x,y) \leqslant \max(d(x,z),d(y,z)).$$

Conversely, if the condition above holds for a given distance d, there exists a dendrogram T such that the corresponding distance is d.

We give the proof of the first part of proposition 1.2.4 as a mean for the reader to become acquainted with some concepts on trees, and we refer to [BG88] for a proof of the second part.

Proof of proposition 1.2.4. Consider a distance d derived from a dendrogram T. For every leaves x and y, let (xy) denote the least common ancestor of x and y. As an example, we represent on the tree T_{ex} the least common ancestor of species A, B, and C in figure 1.14.

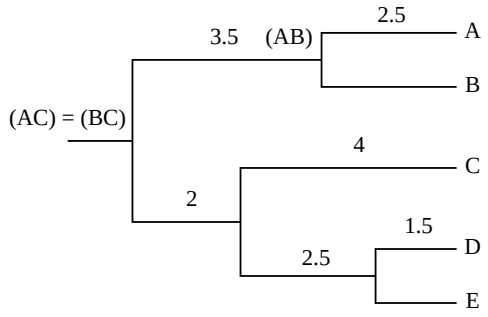


Figure 1.14: Least common ancestors of species A, B and C on tree T_{ex}.

We note that $d(x,y) = 2d(x,(xy))$ for every x and y. Consider three distinct species x, y and z. Two cases arise.

In the first case, there exists t, u and v such that $\{t,u,v\} = \{x,y,z\}$ and (uv) is not ancestral to t. (For instance, in figure 1.14, (AB) is not ancestral to C.) Without loss of generality, one can assume that $t = z$, hence (xy) is not ancestral to z. Since (yz) and (xy) are ancestral to y, one of them is ancestral to the other. Since (xy) is not ancestral to z whereas (yz) is, we deduce that (yz) is ancestral to (xy). Thus, (yz) is ancestral to x, and (yz) is a common ancestor of x and z. Hence, (yz) is ancestral to (xz). Exchanging the roles of x and y, one sees that (xz) is ancestral to (yz), hence $(yz) = (xz)$. This shows that $d(x,z) = 2d(x,(xz)) = 2d(y,(yz)) = d(y,z)$. In other words, the two largest numbers amongst $d(x,y)$, $d(x,z)$ and $d(y,z)$ coincide, and the three-point condition holds.

In the second case, for every t, u and v such that $\{t,u,v\} = \{x,y,z\}$, (uv) is ancestral to t. Then, (xy) is ancestral to z. Since (xy) is also ancestral to x, (xy) is a common ancestor of x and z. Since (xz) is the least common ancestor of x and z, (xy) is ancestral to (xz). Similarly, (xy) is ancestral to (xz). We also know that (yz) is ancestral to x, and that (xz) is ancestral to y. The same arguments show that (yz) is ancestral to (xy) and to (xz), and that (xz) is ancestral to (xy) and to (yz). Finally, $(xy) = (xz) = (yz)$, hence the three-point condition holds. □

Distances on trees in general are characterized by the *four-point condition*.

Proposition 1.2.5 (Four-points condition)**.** *Let T be a (rooted or unrooted) tree with distance d. Then, for every leaves w, x, y and z,*

$$d(w,x) + d(y,z) \leqslant \max\{d(w,y) + d(x,z), d(w,z) + d(x,y)\}.$$

Conversely, if the condition above holds for a distance d, there exists a tree T such that the distance d is derived from T.

We do not provide the proof. However, note that without loss of generality the situation is as in figure 1.15, where the paths from w to x and from y to z do not intersect. This implies that

$$d(w,x) + d(y,z) \leqslant d(w,y) + d(x,z) = d(w,z) + d(x,y).$$

In other words, the two largest sums are equal.

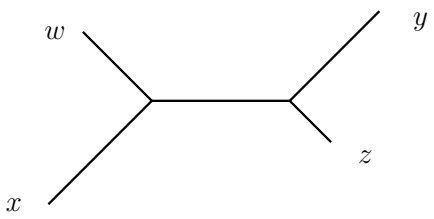


Figure 1.15: Four-point condition.

The three-points and four-points conditions show that it is not always possible to build a tree corresponding to a set of distances. We now describe some distance-based methods to reconstruct phylogenetic trees when this is possible.

Agglomerative algorithms for dendrograms

The basic algorithms in agglomerative approaches are UPGMA or WPGMA (unweighted or weighted pair group method using arithmetic averages) introduced by Sneath and Sokal [SS73]. These algorithms iteratively find pairs of neighbours in

the tree, separate them from the rest of the tree, and reduce the size of the problem by treating the new pair as one unit. Then one recomputes a distance matrix with fewer entries, and one continues with the same approach on the smaller data set. These algorithms assume that the distance matrix satisfies approximately the three-points condition, that is, that the tree one wants to reconstruct is approximately a dendrogram.

Initialization Given n nodes $(S_i)_{1\leqslant i\leqslant n}$, given an input distance matrix Δ with entries δ_{ij}, given n heights $(h_i)_{1\leqslant i\leqslant n}$ initialized to zero,

Step 1 Find clusters i and j such that $i \neq j$ and δ_{ij} is minimal.

Step 2 Define a new height $h_{(ij)} = \delta_{ij}/2$, and create a new node $S_{(ij)}$ with heigh $h_{(ij)}$. Join S_i to S_j at the node $S_{(ij)}$, with the length of branch $S_k S_{(ij)}$ equal to $h_{(ij)} - h_k$ for $k = i, j$.

Step 3 If i and j are the only two entries of Δ, stop and return the tree.

Step 4 Else, build a new distance matrix by removing i and j, and adding (ij), with $\delta_{(ij)k}$ defined as the average of δ_{ik} and δ_{jk} with $k \neq i, j$.

Step 5 Return to step 1 with a distance matrix of a smaller dimension.

Step 4 computes the new distances as the average of two distances that have been previously computed. In WPGMA, the calculation of new distances does not depend on the size of clusters involved, that is,

$$\delta_{k(ij)} = (\delta_{ki} + \delta_{kj})/2.$$

In UPGMA, the average of two distances between clusters i and j depends on their size, that is,

$$\delta_{k(ij)} = (|i|\delta_{ki} + |j|\delta_{kj})/(|i| + |j|).$$

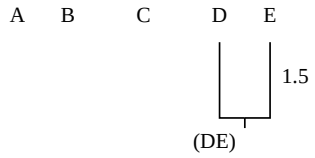
The two algorithms are often confused and some implementations of "UPGMA" correspond in fact to WPGMA.

We now illustrate WPGMA on Δ_{ex}.

Initialization: $h_A = h_B = h_C = h_D = h_E = 0$, $\Delta = \Delta_{ex}$.

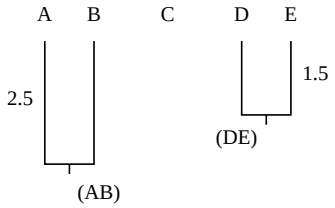
Cycle 1: $\delta_{DE} = 3$ is minimal, $h_{(DE)} = 1.5$,

$$\Delta = \begin{array}{c|ccc} & \text{A} & \text{B} & \text{C} \\ \hline \text{B} & 5 & & \\ \text{C} & 12 & & \\ \text{(DE)} & 12 & 12 & 8 \end{array}$$



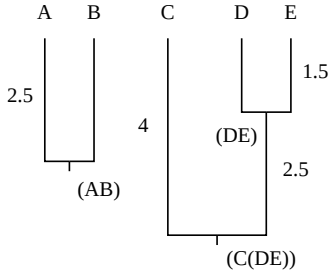
Cycle 2: $\delta_{AB} = 5$ is minimal, $h_{(AB)} = 2.5$,

$$\Delta = \begin{array}{c|cc} & \text{(AB)} & \text{C} \\ \hline \text{C} & 12 & \\ \text{(DE)} & 12 & 8 \end{array}$$

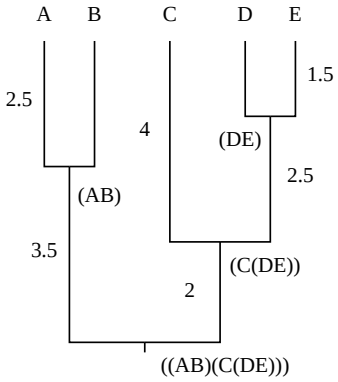


Cycle 3: $\delta_{C(DE)} = 8$ is minimal, $h_{(C(DE))} = 4$,

$$\Delta = \begin{array}{c|c} & \text{(AB)} \\ \hline \text{(C(DE))} & 12 \end{array}$$



Cycle 4: $\delta_{(AB)(C(DE))} = 12$ is minimal, $h_{((AB)(C(DE)))} = 6$, return tree.



The reconstructed tree is similar to the true tree T_{ex}.

If the three-point condition does not hold, the WPGMA method can return erroneous phylogenetic trees. If the four-point condition holds, even approximately,

one uses instead the neighbour joining algorithm (NJ), first introduced by Saitou and Nei [SN87], and modified by Studier and Keppler [SK88].

Neighbor-joining algorithm

The neighbor-joining algorithm is also a neighborliness method, designed to find the shortest tree. This is accomplished by sequentially finding neighbors that minimize the total length among trees in which two of the taxa are clustered together, as in the configuration drawn in figure 1.16 (b). The method starts with a starlike tree T_0^n with n taxa such as the one given in figure 1.16 (a). Before giving the algorithm, we detail some quantities involved.

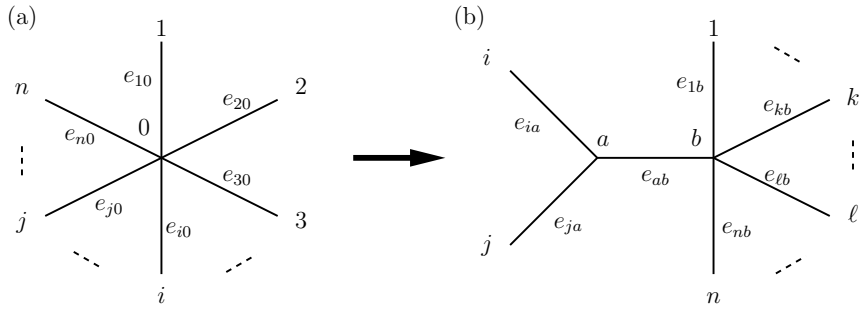


Figure 1.16: Star tree T_0^n with n taxa and tree T_{ij}^n where leaves i and j are clustered.

Let e_{uv} denote the length of the edge uv in the trees T_0^n or T_{ij}^n represented in figure 1.16.

The length $d_{k\ell}$ of the path from the leaf k to the leaf $\ell \neq k$ in the tree T_0^n is $d_{k\ell} = e_{k0} + e_{\ell 0}$. Let $D_0^{(n)}$ denote the total length of the tree T_0^n, that is,

$$D_0^{(n)} = \sum_{k=1}^{n} e_{k0}. \tag{1.2.1}$$

Hence, $(n-1)D_0^{(n)} = R^{(n)}$, with

$$R^{(n)} = \sum_{1 \leqslant k < \ell \leqslant n} d_{k\ell}.$$

Indeed, adding all the distances between the leaves amounts to counting $n-1$ times each branch in the star tree.

In the tree T_{ij}^n, the length $d_{k\ell}$ of the path from leaf k to leaf $\ell \neq k$ is

$$d_{k\ell} = \begin{cases} e_{ka} + e_{ab} + e_{b\ell} & \text{if} \quad k \in \{i,j\} \text{ and } \ell \notin \{i,j\} \\ e_{ia} + e_{ja} & \text{if} \quad \{k,\ell\} = \{i,j\} \\ e_{kb} + e_{\ell b} & \text{if} \quad k \notin \{i,j\} \text{ and } \ell \notin \{i,j\} \end{cases}.$$

Proposition 1.2.6. *Let $D_{ij}^{(n)}$ denote the total length of the tree T_{ij}^n. Then,*

$$2(n-2)\,D_{ij}^{(n)} = 2R^{(n)} - (R_i^{(n)} + R_j^{(n)}) + (n-2)d_{ij},$$

where, for every ℓ,

$$R_\ell^{(n)} = \sum_{k=1}^{n} d_{\ell k}.$$

Proof. One can see on figure 1.16 (b) that

$$D_{ij}^{(n)} = e_{ab} + e_{ia} + e_{ja} + \sum_{k \notin \{i,j\}} e_{kb}. \tag{1.2.2}$$

For every $u \in \{i,j\}$ and every $k \notin \{i,j\}$, one has $d_{uk} = e_{ua} + e_{ab} + e_{kb}$. Summing over k yields

$$\sum_{k \notin \{i,j\}} d_{uk} = (n-2)e_{ua} + (n-2)e_{ab} + \sum_{k \notin \{i,j\}} e_{kb}.$$

Adding d_{ij} to both sides of the equation above yields, for every $u \in \{i,j\}$,

$$R_u^{(n)} = (n-2)e_{ua} + (n-2)e_{ab} + \sum_{k \notin \{i,j\}} e_{kb} + d_{ij},$$

and as a consequence

$$R_i^{(n)} + R_j^{(n)} = nd_{ij} + 2(n-2)e_{ab} + 2\sum_{k \notin \{i,j\}} e_{kb}. \tag{1.2.3}$$

From equations (1.2.2) and (1.2.3), we deduce that

$$2(n-2)D_{ij}^{(n)} = R_i^{(n)} + R_j^{(n)} + (n-4)d_{ij} + 2(n-3)\sum_{k \notin \{i,j\}} e_{kb}. \tag{1.2.4}$$

The subtree formed by the $n-2$ leaves related to the internal node b is a star tree with $n-2$ taxa. From equation (1.2.1), we deduce that

$$(n-3)\sum_{k \notin \{i,j\}} e_{kb} = \sum_{k \notin \{i,j\}} \sum_{\ell \notin \{i,j,k\}} d_{k\ell} = R^{(n)} - (R_i^{(n)} + R_j^{(n)}) + d_{ij}. \tag{1.2.5}$$

Using equation (1.2.5) into equation (1.2.4) yields the result. □

The lengths of the branches from i to a and from j to a in tree T_{ij}^n are such that

$$2(n-2)e_{ia} = (n-2)d_{ij} + R_i^{(n)} - R_j^{(n)}, \quad 2(n-2)e_{ja} = (n-2)d_{ij} + R_j^{(n)} - R_i^{(n)}.$$

For every leaf $k \notin \{i,j\}$,

$$2d_{ka} = d_{ki} + d_{kj} - d_{ij}.$$

We can now provide the algorithm.

Initialization Given $n \geqslant 2$ taxa denoted 1, 2, …, n, given an input distance matrix Δ with entries δ_{ij}, given star tree T_0^n.

Step 1 If $n = 2$, delete 0, join the two leaves with a branch length equal to δ_{12}, and return tree.
Else, compute the $\frac{1}{2}n(n-1)$ values $q_{k\ell}$ as follows

$$q_{k\ell} = (n-2)\delta_{k\ell} - \sum_{m=1}^{n} \delta_{km} - \sum_{m=1}^{n} \delta_{\ell m}.$$

Step 2 Find a pair of taxa i and j such that $q_{ij} \leqslant q_{k\ell}$ for every k and ℓ in $\{1, \ldots, n\}$.

Step 3 Create a node (ij) on the tree that joins the two taxa i and j, and the central node 0 with the lengths ℓ_i and ℓ_j of branches $i(ij)$ and $j(ij)$ equal to

$$\ell_i = \frac{1}{2}\delta_{ij} + \frac{1}{2(n-2)} \left(\sum_{m=1}^{n} \delta_{im} - \sum_{m=1}^{n} \delta_{jm} \right),$$

and

$$\ell_j = \frac{1}{2}\delta_{ij} + \frac{1}{2(n-2)} \left(\sum_{m=1}^{n} \delta_{jm} - \sum_{m=1}^{n} \delta_{im} \right).$$

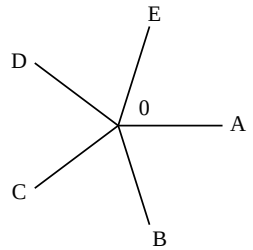
Step 4 Build a new distance matrix by removing i and j, and adding (ij), with $\delta_{(ij)k}$ defined as

$$\delta_{k(ij)} = \frac{1}{2}(\delta_{ki} + \delta_{kj} - \delta_{ij}).$$

Step 5 Return to step 1 with a distance matrix of a smaller size and considering the pair of joined neighbors as a single taxon (ij).

We now illustrate NJ on Δ_{ex}.

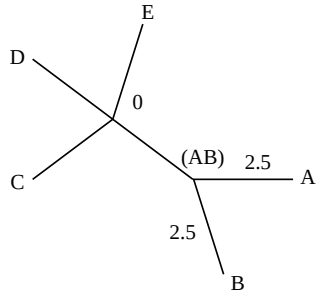
Initialization: $\Delta = \Delta_{ex}$.



Cycle 1: Q is computed, $q_{AB} = -67$ is minimal, $\ell_A = 2.5$, $\ell_B = 2.5$.

$$Q = \begin{array}{c|cccc} & \text{A} & \text{B} & \text{C} & \text{D} \\ \hline \text{B} & -67 & & & \\ \text{C} & -45 & -45 & & \\ \text{D} & -40 & -40 & -51 & \\ \text{E} & -40 & -40 & -51 & -61 \end{array}$$

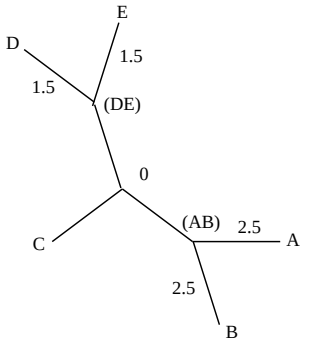
$$\Delta = \begin{array}{c|ccc} & (\text{AB}) & \text{C} & \text{D} \\ \hline \text{C} & 9.5 & & \\ \text{D} & 9.5 & 8 & \\ \text{E} & 9.5 & 8 & 3 \end{array}$$



Cycle 2: Q is computed, $q_{DE} = -35$ is minimal, $\ell_D = 1.5$, $\ell_E = 1.5$.

$$Q = \begin{array}{c|ccc} & (\text{AB}) & \text{C} & \text{D} \\ \hline \text{C} & -35 & & \\ \text{D} & -30 & -30 & \\ \text{E} & -30 & -30 & -35 \end{array}$$

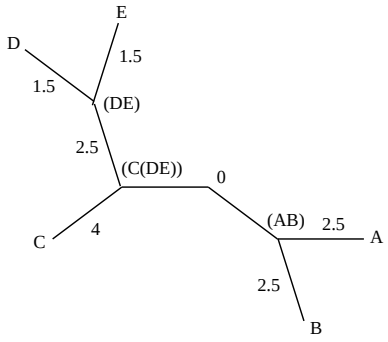
$$\Delta = \begin{array}{c|cc} & (\text{AB}) & \text{C} \\ \hline \text{C} & 9.5 & \\ (\text{DE}) & 8 & 6.5 \end{array}$$



Cycle 3: Q is computed, $q_{C(DE)} = -24$ is minimal, $\ell_C = 4$, $\ell_{(DE)} = 2.5$.

$$Q = \begin{array}{c|cc} & (\text{AB}) & \text{C} \\ \hline \text{C} & -24 & \\ (\text{DE}) & -24 & -24 \end{array}$$

$$\Delta = \begin{array}{c|c} & (\text{AB}) \\ \hline (\text{C(DE)}) & 5.5 \end{array}$$



Cycle 4: $n = 2$, return tree in figure 1.17.

The reconstructed tree is similar to the true tree T_{ex}, ignoring the root.

1.2.3 Discussion

In this chapter, we described two distance algorithms to understand how to reconstruct phylogenetic trees from a distance matrix. We chose to present them because distance algorithms are fast and build trees with thousand of taxas in a few min-

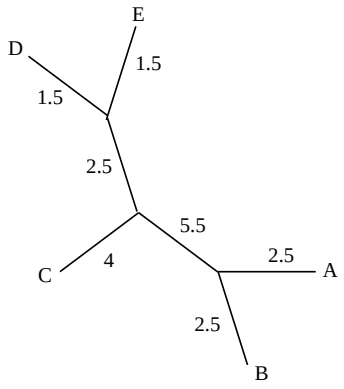


Figure 1.17: Returned tree when one applies NJ to Δ_{ex}

utes whereas other methods (based on maximum likelihood principles for instance) quickly become computationally infeasible.

One should be aware that these two algorithms are now obsolete and have been over-performed ands replaced in practice by some new distance algorithms. However, these are related to NJ, see the discussion at the end of chapter 1 in [Gas05], hence understanding NJ itself is still useful. We did not mention some other questions around these algorithms and, for a deeper review on distance algorithms, we refer to [Gas05] (chapter 1). Finally, we mention that we deal with some Bayesian methods to reconstruct phylogenetic trees in chapter 4.

Chapter 2

Nucleotidic substitution processes

We wrote in chapter 1 that phylogenetics entered the genomic age, and that studies of evolution can now be done at the molecular level. For example, in subsection 1.2.2, we presented some distance-matrix methods of phylogeny reconstruction, but we did not explain how to compute pairwise distances between DNA sequences. Of course, this computation is based on a specific model of nucleotidic substitutions and, more generally, every method of phylogeny reconstruction at the molecular level is based on a substitution model.

In the two first sections of this chapter, we introduce Jukes-Cantor model and Kimura model. We provide estimators of the time of divergence for two contemporary aligned sequences, produced by two copies of the same ancestral unknown sequence evolving independently (this is the homologous case) and for a contemporary sequence and one of its ancestors (this is the ancestral case). In the third section, we present the neighbour dependent substitution models which are the subject of chapter 3.

Some explanations are in order there. We chose to present the classical distance estimation under the Jukes-Cantor model because the strategy used in this very simple case is similar to the one developed in chapter 3 in a more complicated setting. Hence, one can refer to this section while reading chapter 3. About the presentation of Kimura model, an understanding of the maximum likelihood method is not required to understand our results, but, since inferring distances is an important part of this thesis, we felt important to provide a sketch of the use of maximum likelihood principle which is an important tool in phylogeny reconstruction.

The description of the independent models presented below is inspired from [Yan06].

2.1 The Jukes-Cantor model

The Jukes-Cantor model [JC69] is the simplest model of nucleotidic substitution processes in DNA sequences. Relying on the theory of Markov processes, this probabilistic model describes changes between nucleotides. In this section, we present a mathematical description of the model and the main tools to estimate distances between DNA sequences.

2.1.1 Mathematical description of the model

The Jukes-Cantor model (JC) is a continuous-time Markov chain, and our discussion of this model is also meant to introduce some notations and some heuristics useful to understand more complicated models.

Definition 2.1.1. *The alphabet of nucleotides is $\mathscr{A} = \{A,T,C,G\}$, where the letters stand for Adenine, Thymine, Cytosine and Guanine respectively. The set of purines is $R = \{A,G\}$. The set of pyrimidines is $Y = \{T,C\}$.*

In substitution models, a DNA sequence is an element of $\mathscr{A}^N$, where N is a positive integer and stands for the number of nucleotides in one strand of the DNA molecule.

In JC, one assumes that the nucleotidic sites evolve independently from the others, in a similar way. Hence, the evolution of the DNA sequence is ruled by the independent parallel evolutions of N nucleotidic sites. Another assumption of JC is that every substitution occurs at the same rate.

Dynamics of one site

We present now the dynamics of one nucleotidic site in JC.

Definition 2.1.2. *Let $X_i(t)$ denote the random value of the nucleotide at site i and time t. In JC, the process $(X_i(t))_{t\geqslant 0}$ is a Markov process on $\mathscr{A}$ and there exists a parameter $\lambda > 0$ such that the infinitesimal generator $Q = (q_{xy})_{xy}$ is given by the 4×4 matrix of substitution rates*

$$Q = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{array}.$$

Each off-diagonal entry q_{xy} is the rate of substitution of the nucleotide x by the nucleotide y, that is, during a small interval of time $\mathrm{d}t$, x is replaced by y with

probability $q_{xy}\mathrm{d}t$. Each number $-q_{xx}$ can be interpreted as the rate of substitution of the nucleotide x. This means that site i occupied by nucleotide x is modified after an exponentially distributed random time with mean $-1/q_{xx}$, and that, when it is modified, it becomes occupied by nucleotide y with probability $-q_{xy}/q_{xx}$, for each $y \neq x$.

In JC, this means for example that nucleotide C becomes an A, a T or a G with probability $\frac{1}{3}$ each and that the rate of substitution of C is 3λ. The parameter λ rules the overall rate of evolution.

The infinitesimal generator Q fully determines the dynamics of the Markov chain. For instance, let $p_{xy}(t)$ denote the probability that site i is occupied by y at time t given that it is occupied by x at time 0, that is,

$$p_{xy}(t) = \mathbb{P}(X_i(t) = y | X_i(0) = x).$$

Then, the transition-probability matrix over any time t, denoted $P(t) = (p_{xy}(t))$, is determined Q as the unique solution of the initial value problem

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = QP(t), \quad P(0) = \text{Identity},$$

The solution is

$$P(t) = \mathrm{e}^{Qt}.$$

The reader will find a complete presentation of continuous-time Markov chains on discrete spaces in [Nor97]. We now provide the transition-probability matrix for JC.

Proposition 2.1.3. *In JC, for every x and every $y \neq x$ in $\mathscr{A}$, $p_{xy}(t) = \frac{1}{3}p(t)$ and $p_{xx}(t) = 1 - p(t)$, with*

$$p(t) = \frac{3}{4}\left(1 - \mathrm{e}^{-4\lambda t}\right).$$

Proof. Let I denote the 4×4 identity matrix and J the 4×4 matrix whose every coefficient is $\frac{1}{4}$. Then, $IJ = JI = J^2 = J$ and $Q = 4\lambda(J - I)$. Hence,

$$\mathrm{e}^{Qt} = \mathrm{e}^{-4\lambda tI}\mathrm{e}^{4\lambda tJ} = \mathrm{e}^{-4\lambda t}\mathrm{e}^{4\lambda tJ}.$$

For every integer $n \geqslant 1$, $J^n = J$, hence

$$\mathrm{e}^{4\lambda tJ} = \sum_{n \geqslant 0} \frac{(4\lambda t)^n}{n!} J^n = I + (\mathrm{e}^{4\lambda t} - 1)J.$$

This means that $\mathrm{e}^{Qt} = \mathrm{e}^{-4\lambda t}I + (1 - \mathrm{e}^{-4\lambda t})J$, which implies the result. □

The transition-probability matrix coupled with an initial distribution for $X_i(0)$ fully determines the law of $X_i(t)$. Indeed, assume that $X_i(0)$ has the initial distribution $\pi = (\pi_x)_{x \in \mathscr{A}}$, that is

$$\mathbb{P}(X_i(0) = x) = \pi_x,$$

then the law of $X_i(t)$, denoted $\pi(t)$, is given by $\pi(t) = \pi P(t)$, that is, for every nucleotide x

$$\pi_x(t) = \mathbb{P}(X_i(t) = x) = \sum_{y \in \mathscr{A}} \pi_y p_{yx}(t).$$

Dynamics of the DNA sequence

Until now, we have presented the evolution of one site. Let $X_{1:N}(t)$ be a shorthand for $(X_1(t), X_2(t), \ldots, X_N(t))$ and denote the random value of the N nucleotides from site 1 to site N at time t in the DNA sequence.

Definition 2.1.4. *In JC, the process $X_{1:N}(t)$ is Markov, with initial distribution π, where π is a distribution on $\mathscr{A}^N$ and transition kernel $Q^{\otimes N}$. In other words, one has for every $x_{1:N} = (x_1, x_2, \ldots x_N) \in \mathscr{A}^N$*

$$\mathbb{P}(X_{1:N}(t) = x_{1:N}) = \prod_{i=1}^{N} \mathbb{P}(X_i(t) = x_i) = \prod_{i=1}^{N} (\pi \mathrm{e}^{Qt})_{x_i}.$$

Definition 2.1.4 is just a translation of the independence of the N sites and their identically distributions.

When the length N of the sequence is large, the number π_x can be interpreted as the initial proportion of nucleotides x in the sequence, as a consequence of the law of large numbers. Indeed, one has

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{X_i(0) = x\} \xrightarrow[N \to +\infty]{a.s.} \mathbb{E}\left(\mathbf{1}\{X_1(0) = x\}\right) = \pi_x.$$

Similarly, $\pi_x(t)$ can be interpreted as the proportion of nucleotides x at time t in the sequence.

The distribution π^* on $\mathscr{A}$, defined as $\pi^*_x = 1/4$ for every nucleotide x, is *stationary* for the Jukes-Cantor process on one site. This means that if the initial distribution of X_i is π^*, then for every time t the distribution of $X_i(t)$ is π^*. Furthermore, the Jukes-Cantor process is *ergodic*, which means for example that the law of $X_i(t)$ converges to π^* when t tends to infinity, for any initial distribution of X_i.

As a consequence, the proportion of nucleotides x in the DNA sequence when t is large is almost $1/4$, and one can imagine that the sequence is really blended in comparison of the initial sequence, and that pick some information in the sequence may be difficult.

2.1.2 Distance estimation

Since we have presented the Jukes-Cantor model, we now discuss the possibility to calculate pairwise distances between DNA sequences.

One Estimator of the time elapsed between ancestral and present sequences

Consider two DNA sequences of length N. Assume that one sequence is ancestral to the other and that the present sequence has evolved under the Jukes-Cantor model since an unknown time t. This means that every nucleotide site i in the past sequence is ancestral to nucleotide site i in the present sequence. We want to provide an estimator of the elapsed time between these two sequences.

First of all, note that the law of $X_i(t)$ under the Jukes-Cantor model with parameter λ is similar to the law of $X_i(t/2)$ under the Jukes-Cantor model with parameter 2λ. Indeed, in the transition-probability matrix $P(t)$, these quantities appear only in the form of a product λt. Thus, without external information on λ, there is no hope to estimate t in function of λ. Yet, 3λ is the global rate of substitutions per site per unit of time. Given the organism, the nature of DNA, or the area of DNA we look at, maybe some information can be provided on λ, but that is not our purpose here.

In Jukes-Cantor model, the estimator used for the elapsed time is simply based on the proportion of different sites in the two sequences. This is a direct application of the maximum likelihood method detailed for Kimura model in section 2.2.

Precisely, let P_{obs} denote the observed quantity defined as

$$P_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} K_i(t), \quad \text{with} \quad K_i(t) = \mathbf{1}\{X_i(t) \neq X_i(0)\}.$$

The random variables $(K_i(t))_{i=1}^N$ are Bernoulli random variables identically distributed, and their common mean is $p(t)$. Indeed, for every initial distribution π on $\mathscr{A}$, one has

$$\mathbb{P}(X_i(t) = X_i(0)) = \sum_{x\in\mathscr{A}} \pi_x p_{xx}(t) = \sum_{x\in\mathscr{A}} \pi_x (1-p(t)) = 1-p(t).$$

The random variables $(K_i(t))_{i=1}^N$ are also independent, then the law of large numbers and central limit theorem provide

$$P_{\text{obs}} \xrightarrow[N\to+\infty]{a.s.} p(t), \quad \sqrt{N}(P_{\text{obs}} - p(t)) \xrightarrow[N\to+\infty]{d.} \mathscr{N}(0, p(t)(1-p(t))), \qquad (2.1.1)$$

where $\mathscr{N}(\alpha,\sigma^2)$ stands for a normal law with mean α and variance σ^2.

Now we explain how it is possible to compute an estimator for λt from P_{obs}. The function $d \mapsto \frac{3}{4}(1-\mathrm{e}^{-4d})$ is increasing on $[0,+\infty[$ from the value 0 at $d=0$ to the value $3/4$ at $d=+\infty$, and every value p in the interval $[0,3/4[$ corresponds to a unique value d in the interval $[0,+\infty[$ via this function.

This one to one correspondence allows to define the estimator D of the time elapsed.

Definition 2.1.5. *Let D denote the estimator of the time elapsed defined as the solution in d of the equation*

$$\frac{3}{4}(1-\mathrm{e}^{-4d}) = P_{\text{obs}}.$$

In the context of neighbour dependent models which is the content of chapter 3, the estimator of the time elapsed is also defined as the solution of an equation.

Consistency of the estimator and asymptotic confidence interval

Consistency of estimator D is a consequence of the almost convergence of P_{obs} and the continuity of the reciprocal function of $d \mapsto \frac{3}{4}(1-\mathrm{e}^{-4d})$.

Proposition 2.1.6. *The random variable D converges almost surely to λt when N tends to infinity. Hence, D is a consistent estimator of λt.*

Consistency is a nice property for an estimator, but we are also interested in an asymptotic confidence interval for estimator D. Thanks to central limit theorem, we already have an asymptotic confidence interval for $p(t)$, and we now use the delta method [vdV98] to provide an asymptotic confidence interval for t.

Citing [vdV98], the delta method consists of using a Taylor expansion to approximate a random vector of the form $\varphi(T_N)$ by the polynomial $\varphi(\theta)+\varphi'(\theta)(T_N-\theta)+\dots$ in $T_N-\theta$. It is a simple but useful method to deduce the limit law of $\varphi(T_N)-\varphi(\theta)$ from that of $T_N-\theta$.

Proposition 2.1.7 (delta method)**.** *Let $\varphi : I \subset \mathbb{R} \to \mathbb{R}$ be a map defined on an interval I of $\mathbb{R}$ and differentiable at θ. Let (T_N) be random variables taking their values in the domain of φ. Assume that $r_N(T_N-\theta)$ converges in distribution to a random variable T, with $r_N \to +\infty$.*
Then $r_N(\varphi(T_N)-\varphi(\theta))$ converges in distribution to $\varphi'(\theta)T$.

The situation we have with D and P_{obs} is exactly these of proposition 2.1.7. Indeed, let μ denote the reciprocal function of $d \mapsto \frac{3}{4}(1-\mathrm{e}^{-4d})$. Function μ is differentiable on $[0,3/4[$, and

$$D = \mu(P_{\text{obs}}), \quad \lambda t = \mu(p(t)).$$

From convergence in distribution (2.1.1), we deduce

Corollary 2.1.8. *The random variables $\sqrt{N}(D-\lambda t)$ converges in distribution to the centered normal law with variance $\sigma^2(t)$, where*

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

To build a confidence interval for λt from corollary 2.1.8 requires to know the value of $p(t)$ which depends on the quantity λt to be estimated. Slutsky'lemma (see [vdV98]) allows to bypass this difficulty. Indeed, Slutsky's lemma states that if two sequence of random variable $(X_N)_N$ and $(Y_N)_N$ are such that $(X_N)_N$ converges in distribution to a random variable X and $(Y_N)_N$ converges in probability to a constant

c then the sequence $(X_N Y_N)_N$ converges in distribution to the random variable cX. In our situation, we apply this lemma by using the fact that

$$\frac{3-4P_{\text{obs}}}{\sqrt{P_{\text{obs}}(1-P_{\text{obs}})}} \xrightarrow[N\to+\infty]{a.s.} \frac{3-4p(t)}{\sqrt{p(t)(1-p(t))}}.$$

Hence from central limit theorem for P_{obs}, delta method and Slutsky's lemma, one has the final result.

Theorem 2.1.9. *In the Jukes-Cantor model,*

$$(3-4P_{\text{obs}})\sqrt{\frac{N}{P_{\text{obs}}(1-P_{\text{obs}})}}(D-\lambda t) \xrightarrow[N\to+\infty]{d.} \mathcal{N}(0,1).$$

In the context of neighbour dependent models, we use the same strategy than in this section to provide asymptotically Gaussian confidence intervals for the precision of estimation.

At the moment, we have only provided an estimator of the time elapsed between an ancestral sequence and a present one. Remember that in most cases, data are not available for extinct ancestors. We now explain how it is possible to derive an estimator of the time of divergence between two present sequences, that is, a phylogenetic distance between these two DNA sequences

Estimation of the time of divergence between two present sequences

In the case of the Jukes-Cantor model, we provide an estimator for the time of divergence between two present sequences directly from the estimator of the time elapsed between an ancestral sequence and a present one. However, it is necessary to assume that the ancestral sequence is at stationarity.

The Jukes-Cantor model is a time-reversible Markov chain. Reversibility means that the dynamics will look the same whether time runs forward or backward. To check if a Markov process is time-reversible, one can use the generalized Kolmogorov criterion (Kendall [Ken59]), that is, one needs to check that for each loop in the state space, the product of the rates is the same whatever direction in the loop is chosen.

Proposition 2.1.10 (Generalized Kolomogorov criterion)**.** *A stationary Markov process with infinitesimal generator Q and state space $\mathscr{S}$ is reversible if and only if the entries of Q satisfy*

$$q_{x_1x_2}q_{x_2x_3}\cdots q_{x_{n-1}x_n}q_{x_nx_1} = q_{x_1x_n}q_{x_nx_{n-1}}\cdots q_{x_3x_2}q_{x_2x_1},$$

for any elements $x_1, x_2, \ldots, x_n \in \mathscr{S}$.

As a result, given two sequences at stationarity, the probability of data in a state is the same whether one sequence is ancestral to the other or both are descendants of an ancestral sequence at stationarity. Roughly speaking, for every x and y that belong to $\mathscr{A}$, going from a x at time t to 0 then back to a y at time t on another branch, is equivalent to going from a x at time 0 to a y at time $2t$, as illustrated in figure 2.1.

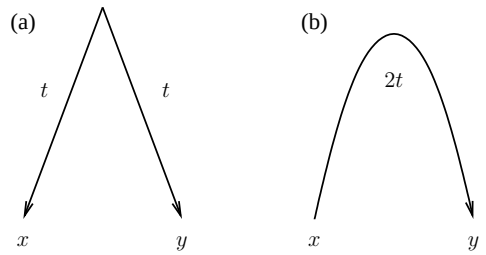


Figure 2.1: A tree for two sequences showing the observed nucleotides x and y at one site. (a) Two sequences diverged from a common ancestor. (b) Sequence 1 is ancestral to sequence 2.

To understand why stationarity is necessary, imagine the following case. Let the time run forward largely from the ancestral sequence to sequence1. Then, sequence 1 is almost at stationarity. As a consequence, the proportion of each nucleotide in sequence 1 is almost $1/4$. Now, let the time run backward from sequence 1 to the ancestral sequence, as the time is large, the ancestral sequence is also at stationarity. Stationarity is necessary for time-reversible process.

Let $X_{1:N}^k(t)$ denote for every $k \in \{1,2\}$, the random value of the N nucleotides from site 1 to site N at time t in the DNA sequence k.

Let $\tilde{P}_{\text{obs}}$ denote the observed quantity defined as

$$\tilde{P}_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} \tilde{K}_i(t), \quad \text{with} \quad \tilde{K}_i(t) = \mathbf{1}\{X_i^1(t) \neq X_i^2(t)\}.$$

Definition 2.1.11. *Let $\tilde{D}$ denote the estimator of the time elapsed defined as the solution in d of the equation*

$$\frac{3}{4}(1 - \mathrm{e}^{-8d}) = \tilde{P}_{\text{obs}}.$$

The main result of this section is the following.

Theorem 2.1.12. *Assume that two present DNA sequences have diverged under the Jukes-Cantor model with parameter λ from a common ancestral DNA sequence at equilibrium since a time t.*

Then, $\tilde{D}$ is a consistent estimator of λt, and

$$(6-8\tilde{P}_{\text{obs}})\sqrt{\frac{N}{\tilde{P}_{\text{obs}}(1-\tilde{P}_{\text{obs}})}}(\tilde{D}-\lambda t) \xrightarrow[N\to+\infty]{d.} \mathcal{N}(0,1).$$

2.2 Other independent models of evolution

In section 2.1, we have presented the simplest nucleotide substitution model. We now discuss probabilistic models where the independence between nucleotide sites is still assumed, but where some constrains are placed on substitution rates.

2.2.1 Model of Kimura

In subsection 1.1.3, we classified substitutions in two types: transitions, which exchange a purine for a purine ($A \leftrightarrow G$) or a pyrimidine for a pyrimidine ($C \leftrightarrow G$), and transversions which exchange a purine for a pyrimidine or a pyrimidine for a purine ($C/T \leftrightarrow A/G$). This classification comes from one observation on real data: transitions often occur at higher rate than transversions. Thus, Kimura [Kim80] proposed a probabilistic model derived from the Jukes-Cantor one which takes into account the difference between transition and transversion rates.

Mathematical description of Kimura's model

Kimura's model is as the Jukes-Cantor model a continuous-time Markov process.

Definition 2.2.1. *In Kimura's model, the process $(X_i(t))_{t\geqslant 0}$ is a Markov process on $\mathscr{A}$ whose infinitesimal generator Q is given by the 4×4 matrix of substitution rates*

$$Q = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} -(\alpha+2\beta) & \beta & \beta & \alpha \\ \beta & -(\alpha+2\beta) & \alpha & \beta \\ \lambda & \alpha & -(\alpha+2\beta) & \beta \\ \alpha & \beta & \beta & -(\alpha+2\beta) \end{pmatrix} \end{array}.$$

where α and β are positive parameters.

Parameter α represents the rate of transitions, whereas parameter β represents the rate of transversions.

The dynamics of one site is given by the transition-probability matrix whose entries are defined as below.

Proposition 2.2.2. *In Kimura's model, for every couple of nucleotides (x,y), one has*

$$p_{xy}(t) = \begin{cases} p_0(t) & \text{if} \quad x = y \\ p_1(t) & \text{if} \quad \{x,y\} = R \text{ or } \{x,y\} = Y \\ p_2(t) & \text{else if} \end{cases},$$

with

$$\begin{aligned} 4p_0(t) &= 1 + \mathrm{e}^{-4\beta t} + 2\mathrm{e}^{-2(\alpha+2\beta)t}, \\ 4p_1(t) &= 1 + \mathrm{e}^{-4\beta t} - 2\mathrm{e}^{-2(\alpha+2\beta)t}, \\ 4p_2(t) &= 1 - \mathrm{e}^{-4\beta t}. \end{aligned}$$

Proof. Note that $Q = -2(\alpha+\beta)I + \alpha J_1 + \beta J_2$ where I stands for the identity matrix and J_1 and J_2 are defined as

$$J_1 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad J_2 = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix}.$$

Matrices J_1, J_2 and J, where J stands for the matrix whose every entry is equal to 1, satisfies the following equations

$$J = J_1 + J_2, \quad \text{and} \quad J_1 J_2 = J_2 J_1 = 2J_2$$

Hence, one has

$$\mathrm{e}^{Qt} = \mathrm{e}^{-2(\alpha+\beta)tI} \mathrm{e}^{\alpha t J_1} \mathrm{e}^{\beta t J_2} = \mathrm{e}^{-2(\alpha+\beta)t} \mathrm{e}^{(\alpha-\beta)tJ_1} \mathrm{e}^{\beta t J}.$$

As for every positive integer n, $J^n = 4^{n-1}J$ and $J_1^n = 2^{n-1}J_1$, one has

$$\mathrm{e}^{(\alpha-\beta)tJ_1} = I + \frac{\mathrm{e}^{2(\alpha-\beta)t} - 1}{2} J_1 \quad \text{and} \quad \mathrm{e}^{\beta t J} = I + \frac{\mathrm{e}^{4\beta t} - 1}{4} J.$$

Noting that $JJ_1 = 2J$, and multiplying the last two equalities yields

$$\mathrm{e}^{(\alpha-\beta)tJ_1} \mathrm{e}^{\beta t J} = I + \frac{\mathrm{e}^{2(\alpha+\beta)t} + \mathrm{e}^{2(\alpha-\beta)t} - 2}{4} J_1 + \frac{\mathrm{e}^{2(\alpha+\beta)t} - \mathrm{e}^{2(\alpha-\beta)t}}{4} J_2.$$

Finally,

$$\mathrm{e}^{Qt} = \mathrm{e}^{-2(\alpha+\beta)t} I + \frac{1 + \mathrm{e}^{-4\beta t} - 2\mathrm{e}^{-2(\alpha+\beta)t}}{4} J_1 + \frac{1 - \mathrm{e}^{-4\beta t}}{4} J_2,$$

and this achieves the proof □

As for the Jukes-Cantor model, the uniform distribution π^* on $\mathscr{A}$ is stationary for the Kimura's model. The global rate of substitution per time and per site is $\alpha + 2\beta$.

The dynamics of the whole sequence is, as the Jukes-Cantor model, the independent dynamics of the N sites. We now explain the method to provide phylogenetic distances in Kimura's model.

2.2.2 Distance estimation

As in the Jukes-Cantor model, it is impossible to estimate t without information on α and β. For Kimura's model, it is convenient to use parameter $d = (\alpha + 2\beta)t$, which will be the distance between the two sequences, and the transition/transversion rate ratio $\kappa = \alpha/\beta$. Using these notations yields

$$\begin{aligned}
4p_0(t) &= 1 + \mathrm{e}^{-4d/(\kappa+2)} + 2\mathrm{e}^{-2d(\kappa+1)/(\kappa+2)}, \\
4p_1(t) &= 1 + \mathrm{e}^{-4d/(\kappa+2)} - 2\mathrm{e}^{-2d(\kappa+1)/(\kappa+2)}, \\
4p_2(t) &= 1 - \mathrm{e}^{-4d/(\kappa+2)}.
\end{aligned}$$

We now discuss the maximum likelihood method for estimating sequence distances in Kimura's model. Maximum likelihood is a general methodology for estimating parameters in a model.

For a fixed set of data and underlying probability model, maximum likelihood picks the values of the model parameters that make the data "more likely" than any other values of the parameters would make them.

Suppose there is a sample x_1, x_2, ..., x_n of n independent observations, drawn from an unknown probability density (or probability mass) f_0. We assume that the function f_0 belongs to a certain family of distributions $\{f(\cdot|\theta); \theta \in \Theta\}$, called the parametric model, so that f_0 corresponds to $\theta = \theta_0$, which is called the true value of the parameter. The idea behind the method of maximum likelihood is to look at the joint density function $f(x_1, x_2, \ldots, x_n|\theta)$ at a different angle. Let the observed values $x_1, x_2, \ldots, x_n$ be fixed "parameters" of this function, wheras the value of θ is allowed to vary freely. From this point of view this function is called the likelihood and denoted

$$L(\theta|x_1, \ldots, x_n) = f(x_1, x_2, \ldots, x_n|\theta).$$

In practice it is always more convenient to work with the logarithm of the likelihood function, called the log-likelihood:

$$\ell(\theta|x_1, \ldots, x_n) = \log[L(\theta|x_1, \ldots, x_n)].$$

The method of maximum likelihood estimates θ_0 by finding the value of θ that maximizes $\ell(\theta|x)$. For detailed properties of maximum likelihood estimator, the reader will consult [vdV98].

In Kimura's model, the data are the number of sites with transitional NS_{obs} and transversional NV_{obs} differences, where

$$\begin{aligned}
S_{\text{obs}} &= \frac{1}{N}\sum_{i=1}^{N} \left(\mathbf{1}\{\{X_i(0), X_i(t)\} = R\} + \mathbf{1}\{\{X_i(0), X_i(t)\} = Y\}\right), \\
V_{\text{obs}} &= \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{X_i(t) \neq X_i(0)\} - S_{\text{obs}}.
\end{aligned}$$

The number of constant sites is $N - NS_{\text{obs}} - NV_{\text{obs}}$.

The probability to observe a transition is given by $p_1(t)$, the probability to observe a transversion is $p_2(t)$, and the probability is $p_0(t)$ for any constant site. Hence, the likelihood function is

$$L(d,\kappa;n_S,n_V) = \binom{N}{n_S\, n_V} p_1(t)^{n_S} p_2(t)^{n_V} p_0(t)^{N-n_S-n_V}.$$

The log-likelihood is defined, up to an additive term not depending on t, as

$$\ell(d,\kappa;n_S,n_V) = n_S \log(p_1(t)) + n_V \log(p_2(t)) + (N - n_S - n_V)\log(p_0(t)).$$

Maximum likelihood estimator can be derived from the equations $\partial\ell/\partial d = 0$ and $\partial\ell/\partial\kappa = 0$. The solution can be shown to be

Proposition 2.2.3. *Let D and K denote estimators defined as*

$$\begin{aligned} D &= -\frac{1}{4}\log(1-2V_{\text{obs}}) - \frac{1}{2}\log(1-2S_{\text{obs}}-V_{\text{obs}}), \\ K &= 2\frac{\log(1-2S_{\text{obs}}-V_{\text{obs}})}{\log(1-2V_{\text{obs}})} - 1. \end{aligned}$$

Then D and K are consistent estimators of $(\alpha+2\beta)t$ *and* α/β.

It is also possible to apply delta method and Slutsky's lemma to D. Hence, one has the following result.

Theorem 2.2.4. *In Kimura's model*

$$\sqrt{\frac{N}{\sigma^2_{\text{obs}}}}(D - (\alpha+2\beta)t) \xrightarrow[N\to+\infty]{d.} \mathcal{N}(0,1),$$

where

$$\sigma^2_{\text{obs}} = A^2_{\text{obs}} S_{\text{obs}} + B^2_{\text{obs}} V_{\text{obs}} - (A_{\text{obs}} S_{\text{obs}} + B_{\text{obs}} V_{\text{obs}})^2,$$

with

$$\begin{aligned} A_{\text{obs}} &= (1-2S_{\text{obs}}-V_{\text{obs}})^{-1}, \\ B_{\text{obs}} &= [(1-2S_{\text{obs}}-V_{\text{obs}})^{-1} + (1-2V_{\text{obs}})^{-1}]/2. \end{aligned}$$

Kimura's model is also time-reversible, and it is possible to derive a distance between two DNA sequences having diverged from a common ancestral DNA sequence at equilibrium as in the Jukes-Cantor model.

2.2.3 Other models

We have presented models of Jukes-Cantor and Kimura. These two models assume that the sites evolve independently from the others, and have the uniform distribution on $\mathscr{A}$ as stationary distribution. In other models of DNA evolution, such as [Fel81], [HKY85] and [TN93], one still assumes the independence between sites but one places different constrains on the rates of substitution. As a consequence, the stationary distribution may be different from the uniform distribution, but with no interaction between the sites, the nucleotide attached to any given site converges in distribution to the stationary measure of the Markov chain described by the matrix of the rates and, at equilibrium, the sites are independent.

There also exists the possibility to modify the rate of substitution for any site, assuming for example that this rate is a random variable drawn from a statistical distribution. But there again, it is possible to derive distances without too much difficulty. The real difficulty is the introduction of substitution rates which depend on the nature of closed neighbours of sites.

2.3 About neighbour dependent substitution processes

As we have seen in subsection 1.1.3, it is well known that the nucleotides in the immediate neighbourhood of a site can affect drastically the substitution rates at this site. For instance, in the genomes of vertebrates, the increased substitutions of cytosine by thymine and of guanine by adenine in CpG dinucleotides are often quite noticeable. The chemical reasons of this CpG-methylation-deamination process are also well known and one can guess that, at equilibrium, the number of CpG is decreased while the number of TpG and CpA is increased when one adds high rates of CpG substitutions.

The need to incorporate an influence of the neighbourhood into more realistic models of nucleotide substitutions seems widely acknowledged. That is the reason why Duret and Galtier introduced and analysed a model in [DG00], which we call Tamura + CpG, that adds to Tamura's rates of substitution the availability of substitutions $CG \to CA$ and $CG \to TG$, both at the additional rate $\rho \geqslant 0$.

However, the exact consequences of the introduction of such neighbour-dependent substitution processes remained virtually unknown, at least up to their knowledge, on a theoretical ground. To understand why, note that the distribution of the nucleotide at site i at a given time depends a priori on the values at previous times of the dinucleotides at sites $(i, i-1)$ and $(i, i+1)$, whose joint distributions, in turn, may depend on the values of some trinucleotides, and so on. Hence, one is faced with infinite-dimensional linear systems, which are difficult to solve.

To evade the curse of recursive calls to the frequencies of longer and longer words, Duret and Galtier used as approximate frequencies (xyz) of the trinucleotides the

values

$$(xyz) \approx (xy)(yz)/(y).$$

Their approximations enabled them to capture some features of the behavior of the true model, but it was not mathematically founded.

Entering in the field of interacting particle systems [Lig85], Bérard, Gouéré and Piau [BGP08] introduced a wide extension of the Tamura + CpG model of neighbour-dependent substitution processes. Even if this field contains more difficulties than the field of finite Markov chains, they showed that these models are solvable. For example, they proved that the frequencies of polynucleotides at equilibrium solve explicit finite-size linear systems.

We now describe these models and their properties.

2.3.1 Jukes-Cantor model with CpG influence

Recall that DNA sequences are encoded by the alphabet $\mathscr{A} = \{A, T, C, G\}$, where the letters stand for Adenine, Thymine, Cytosine and Guanine respectively.

In independent evolution models, DNA sequences are encoded as elements of $\mathscr{A}^N$, where N is a positive integer. In the Jukes-Cantor model with CpG influence (JC+CpG), DNA sequences are encoded as elements of $\mathscr{A}^{\mathbb{Z}}$ where $\mathbb{Z}$ is the set of integers, and as a consequence bi-infinite.

Heuristics of the mechanisms

The probabilistic JC+CpG model is a continuous-time Markov chain on $\mathscr{A}^{\mathbb{Z}}$, where the sequence evolves under the combined effect of two superimposed mechanisms.

The first mechanism is an independent evolution of the sites as in the usual Jukes-Cantor model with parameter 1. Hence it is characterized by a 4×4 matrix of substitution rates, each rate being the mean number of substitutions per unit of time. Hence, the rate of the substitutions of x by y is set to 1, for every nucleotides x and y in $\mathscr{A}$.

A second mechanism is superimposed, which describes the substitutions due to the influence of the neighborhood: the most noticeable case is based on experimentally observed CpG-methylation-deamination processes, whose biochemical causes are well known. Hence we assume that the substitution rates of cytosine by thymine and of guanine by adenine in CpG dinucleotides are both increased by an additional nonnegative rate r.

This means for example that any C site whose right neighbour is not occupied by a G, changes at global rate 3, hence after an exponentially distributed random time with mean $1/3$, as drawn for the C site in $(N-1)$th position on figure 2.2, and when it does, it becomes an A, a G or a T with probability $1/3$ each. On the contrary, any

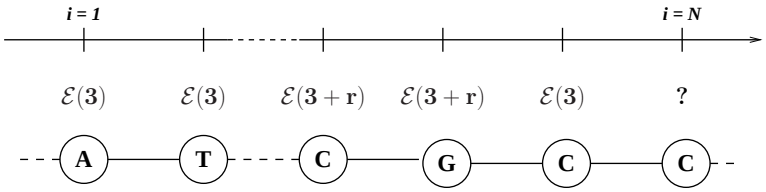


Figure 2.2: A piece of DNA sequence under JC+CpG model with the exponentially distributed random time drawn over each site

C site whose right neighbour is occupied by a G, changes at global rate $s = 3 + r$, hence after an exponentially distributed random time with mean $1/s$, as drawn for the C site in $(N-3)$th position on figure 2.2, and when it does, it becomes an A, a G or a T with unequal probabilities $1/s$, $1/s$, and $(1+r)/s$ respectively. Note that for the C site in Nth position on figure 2.2, one has to look at its right neighbour to know in which case one is.

The case $r = 0$ corresponds to the usual Jukes-Cantor model. As soon as $r \neq 0$, the evolution of a site is not independent of the rest of the sequence. Hence the evolution of the complete sequence is Markovian (on a huge state space), but not the evolution of a given site, nor of any given finite set of sites.

Main properties

We work on the space $\mathscr{A}^{\mathbb{Z}}$ with the topology product and the cylindric σ-algebra defined as the smallest σ-algebra such that every projection on $\mathscr{A}^{\mathbb{Z}}$ is measurable.

We now recall some results of [BGP08].

Theorem 2.3.1 (Bérard, Gouéré and Piau [BGP08])**.** *For every probability measure ν on $\mathscr{A}^{\mathbb{Z}}$, there exists a unique Markov process $(X(t))_{t\geqslant 0}$ on $\mathscr{A}^{\mathbb{Z}}$, with initial distribution ν, associated to the transition rates above.*

Thus, for every time t, $X(t)$ describes the whole sequence and, for every i in $\mathbb{Z}$, the ith coordinate $X_i(t)$ of $X(t)$ is the random value of the nucleotide at site i and time t.

Theorem 2.3.2 (Bérard, Gouéré and Piau [BGP08])**.** *The process $(X(t))_{t\geqslant 0}$ is ergodic, its unique stationary distribution π on $\mathscr{A}^{\mathbb{Z}}$ is invariant and ergodic with respect to the translations of $\mathbb{Z}$, and π puts a positive mass on every finite word $w = (w_i)_{0\leqslant i\leqslant \ell}$ written in the alphabet $\mathscr{A}$.*

The notation $\pi(w)$ is abusive because π is a measure on $\mathscr{A}^{\mathbb{Z}}$ but it is a shorthand for $\pi(\Pi_{0,\ell}^{-1}(\{w\}))$, where $\Pi_{0,\ell}$ is such that for every $x \in \mathscr{A}^{\mathbb{Z}}$, $\Pi_{0,\ell}(x) = (x_i)_{0\leqslant i\leqslant \ell}$.

Furthermore, for every position i in $\mathbb{Z}$, $\mathbb{P}_\nu(X_{i:i+\ell}(t) = w)$ converges to $\pi(w)$ when $t \to +\infty$, where $\mathbb{P}_\nu$ stands for the probability under the initial measure ν. Here

and later on, for every indices i and j in $\mathbb{Z}$ with $i \leqslant j$ and every symbol S, the shorthand $S_{i:j}$ denotes $(S_k)_{i\leqslant k\leqslant j}$. Finally, if ξ in $\mathscr{A}^{\mathbb{Z}}$ is distributed according to π, the empirical frequencies of any word w in ξ, observed along any increasing sequence of intervals of $\mathbb{Z}$, almost surely converge to $\pi(w)$.

All of the above properties stem from the following representation of the distribution π.

Theorem 2.3.3 (Bérard, Gouéré and Piau [BGP08])**.** *There exists an i.i.d. sequence $(\xi_i)_{i\in\mathbb{Z}}$ of Poisson processes, and a measurable map Ψ with values in $\mathscr{A}$, such that if one sets*

$$\Xi_i = \Psi(\xi_{i-1}, \xi_i, \xi_{i+1})$$

for every site i in $\mathbb{Z}$, then the distribution of $(\Xi_i)_{i\in\mathbb{Z}}$ is π.

In particular, any collections $(\Xi_i)_{i\in I}$ and $(\Xi_i)_{i\in J}$ are independent as soon as the subsets I and J of $\mathbb{Z}$ are such that $|i-j| \geqslant 3$ for every sites i in I and j in J. We call this property 2-dependence.

2.3.2 Class of neighbour dependent substitution models

We have presented the JC+CpG model, but this model is just the simplest neighbour-dependent substitution process of the class of models, called RN+YpR introduced in [BGP08]. We briefly introduce this class now.

Firstly, RN stands for Rzhetsky-Nei and means that the 4×4 matrix of substitution rates which characterize the independent evolution of the sites must satisfy 4 equalities, summarized as follows: for every pair of nucleotides x and $y \neq x$, the substitution rate from x to y may depend on x but only through the fact that x is a purine (A or G, symbol R) or a pyrimidine (C or T, symbol Y). For instance, the substitution rates from C to A and from T to A must coincide, likewise for the substitution rates from A to C and from G to C, from C to G and from T to G, and finally from A to T and from G to T. The 4 remaining rates, corresponding to purine-purine substitutions and to pyrimidine-pyrimidine substitutions, are free.

Secondly, the influence mechanism is called YpR, which stands for the fact that one allows any specific substitution rates between any two YpR dinucleotides (CG, CA, TG and TA) which differ by one position only, for a total of 8 independent parameters. The Jukes-Cantor model with CpG effect is the simplest non trivial one: the only YpR substitutions with positive rate are $CG \to CA$ and $CG \to TG$, and both happen at the same rate.

Recall that Y denote the set of pyrimidines defined as $Y = \{T, C\}$, and R the set of purines defined as $= \{A, G\}$.

The 4×4 matrix of substitution rates which characterize the independent evolution

of the sites in RN model is given by

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & v_T & v_C & w_G \\ v_A & \cdot & w_C & v_G \\ v_A & w_T & \cdot & v_G \\ w_A & v_T & v_C & \cdot \end{pmatrix} \end{array}.$$

The influence mechanism called YpR adds specific rates of substitutions from each YpR dinucleotide as follows.

- Every dinucleotide CG moves to CA at rate r_A^C and to TG at rate r_T^G.
- Every dinucleotide TA moves to CA at rate r_C^A and to TG at rate r_G^T.
- Every dinucleotide CA moves to CG at rate r_G^C and to TA at rate r_T^A.
- Every dinucleotide TG moves to CG at rate r_C^G and to TA at rate r_A^T.

Under a non-degeneracy condition (always satisfied if the rates are non negative), theorems 2.3.1, 2.3.2 and2.3.3 occur.

In chapter 3, we show how to compute consistent estimators and asymptotic confidence intervals for the evolutionary time between DNA sequences in these evolution models.

Chapter 3

Toward phylogenetic distances for RN + YpR models

In this chapter we consider models of nucleotidic substitution processes where the rate of substitution at a given site depends on the state of the neighbours of the site. We first estimate the time elapsed between an ancestral sequence at stationarity and a present sequence. Second, assuming that two sequences are issued from a common ancestral sequence at stationarity, we estimate the time since divergence. In the simplest nontrivial case, the Jukes-Cantor model with CpG influence, we provide and justify mathematically consistent estimators in these two settings. We also provide asymptotic confidence intervals, valid for nucleotidic sequences of finite length, and we compute explicit formulas for the estimators and for their confidence intervals. In the general case of an RN model with YpR influence, we extend these results under a proviso, namely that the equation defining the estimator has a unique solution.

Introduction

A crucial step in the computation of phylogenetic trees based on aligned DNA sequences is the estimation of the evolutionary times between these sequences. In most phylogenetic algorithms based on stochastic substitution models, one assumes that each site evolves independently from the others and, in general, according to a given Markovian kernel. This assumption is mainly due to the difficulty to work without the assumption of independence. To understand why, note that, as soon as the rates of substitutions of th the distribution of the nucleotide at site i at a given time depends a priori on the values at previous times of the dinucleotides at sites $i-1$ and $i+1$, whose joint distributions, in turn, may depend on the values of some trinucleotides, and so on. Hence, one is faced with infinite-dimensional linear systems, which are generically hard to solve. Besides, the magnitude of the

effect of the neighbours on the substitution rates can be large. Since some neighbour influences are well documented in the literature, and caused by well known biological mechanisms, it seems necessary to take into account the neighbour influences in substitution models. To wit, a class of mathematical models with neighbour influences was recently introduced by biologists, see [GGG96], and studied mathematically, see [BGP08].

The goal of the present chapter is to show that one can compute consistent estimators of the distances between DNA sequences whose evolution is ruled by models with influence in a specific class of models.

We completely describe the construction in the simplest non trivial case, the Jukes-Cantor model with (symmetric) CpG influence and we explain in sections 3.8 how to extend our construction to every model in the class.

In section 3.1, we describe the Jukes-Cantor model with CpG influence, the simplest one of the class of manageable models introduced in [BGP08], and its main properties. In section 3.2, we summarize our main results on the estimation of the elapsed time between an old DNA sequence and a present one, and on the time since two present DNA sequences issued from the same ancestral sequence diverged. Section 3.8 contains the extension of the results of section 3.2. In the other sections we prove our results. At the end of section 3.2, we detail the plan of the rest of the chapter.

3.1 Models with influence

We first describe the Jukes-Cantor model with CpG influence to which the results of this chapter apply. Then, we mention its main mathematical properties, already established in [BGP08], and we introduce some notations.

Recall that DNA sequences are encoded by the alphabet $\mathscr{A} = \{A, T, C, G\}$, where the letters stand for Adenine, Thymine, Cytosine and Guanine respectively. Thus, bi-infinite DNA sequences are encoded as elements of $\mathscr{A}^{\mathbb{Z}}$ where $\mathbb{Z}$ is the set of integers.

3.1.1 Jukes-Cantor model with CpG influence (JC+CpG)

In most models of DNA evolution, one assumes that each site evolves independently from the others and follows a given Markovian kernel, see [JC69], [Kim80], [Fel81] and [HKY85] for instance. Even in codon evolution models, see [JTT92], one often assumes that different codons evolve independently, with however some exceptions such as [JP00]. On the other hand, it is a well known experimental fact, see [DG00] by example, that the nature of the close neighbours of a site can modify, notably in some cases, the substitution rates observed at this site. To take account of these observations, we consider models, in continuous time, where the

sequence evolves under the combined effect of two superimposed mechanisms.

The first mechanism is an independent evolution of the sites as in the usual models. Hence it is characterized by a 4×4 matrix of substitution rates, each rate being the mean number of substitutions per unit of time. The simplest case is the Jukes-Cantor model, where each substitution happens at the same rate. Hence, the rate of the substitutions of x by y is set to 1, for every nucleotides x and y in $\mathscr{A}$.

A second mechanism is superimposed, which describes the substitutions due to the influence of the neighborhood: the most noticeable case is based on experimentally observed CpG-methylation-deamination processes, whose biochemical causes are well known. Hence we assume that the substitution rates of cytosine by thymine and of guanine by adenine in CpG dinucleotides are both increased by an additional nonnegative rate r.

This means for example that any C site whose right neighbour is not occupied by a G, changes at global rate 3, hence after an exponentially distributed random time with mean $1/3$, and when it does, it becomes an A, a G or a T with probability $1/3$ each. On the contrary, any C site whose right neighbour is occupied by a G, changes at global rate $s = 3 + r$, hence after an exponentially distributed random time with mean $1/s$, and when it does, it becomes an A, a G or a T with unequal probabilities $1/s$, $1/s$, and $(1+r)/s$ respectively.

The case $r = 0$ corresponds to the usual Jukes-Cantor model. As soon as $r \neq 0$, the evolution of a site is not independent of the rest of the sequence. Hence the evolution of the complete sequence is Markovian (on a huge state space), but not the evolution of a given site, nor of any given finite set of sites.

Recall from [BGP08] that the relevant class of models, called RN+YpR, is in fact larger than just described.

As already mentioned, the results of this chapter about Jukes-Cantor models with CpG influence (hereafter denoted JC+CpG) are adapted to every RN model with YpR influence (hereafter denoted RN+YpR) in section 3.8.

3.1.2 Main properties

We work on the space $\mathscr{A}^{\mathbb{Z}}$ with the topology product and the cylindric σ-algebra defined as the smallest σ-algebra such that every projection on $\mathscr{A}^{\mathbb{Z}}$ is measurable.

We now recall some results of [BGP08], valid for every RN+YpR model. First, for every probability measure ν on $\mathscr{A}^{\mathbb{Z}}$, there exists a unique Markov process $(X(t))_{t\geqslant 0}$ on $\mathscr{A}^{\mathbb{Z}}$, with initial distribution ν, associated to the transition rates above. Thus, for every time t, $X(t)$ describes the whole sequence and, for every i in $\mathbb{Z}$, the ith coordinate $X_i(t)$ of $X(t)$ is the random value of the nucleotide at site i and time t. Under a non-degenaracy condition on the rates of the model, the process $(X(t))_{t\geqslant 0}$ is ergodic, its unique stationary distribution π on $\mathscr{A}^{\mathbb{Z}}$ is invariant and ergodic with respect to the translations of $\mathbb{Z}$, and π puts a positive mass on every finite word

$w = (w_i)_{0 \leqslant i \leqslant \ell}$ written in the alphabet $\mathscr{A}$. The notation $\pi(w)$ is abusive because π is a measure on $\mathscr{A}^{\mathbb{Z}}$ but it is a shorthand for $\pi(\Pi_{0,\ell}^{-1}(\{w\}))$, where $\Pi_{0,\ell}$ is such that for every $x \in \mathscr{A}^{\mathbb{Z}}$, $\Pi_{0,\ell}(x) = (x_i)_{0 \leqslant i \leqslant \ell}$.

Furthermore, for every position i in $\mathbb{Z}$, $\mathbb{P}_\nu(X_{i:i+\ell}(t) = w)$ converges to $\pi(w)$ when $t \to +\infty$, where $\mathbb{P}_\nu$ stands for the probability under the initial measure ν. Here and later on, for every indices i and j in $\mathbb{Z}$ with $i \leqslant j$ and every symbol S, the shorthand $S_{i:j}$ denotes $(S_k)_{i \leqslant k \leqslant j}$. Finally, if ξ in $\mathscr{A}^{\mathbb{Z}}$ is distributed according to π, the empirical frequencies of any word w in ξ, observed along any increasing sequence of intervals of $\mathbb{Z}$, almost surely converge to $\pi(w)$.

All of the above properties stem from the following representation of the distribution π. There exists an i.i.d. sequence $(\xi_i)_{i \in \mathbb{Z}}$ of Poisson processes, and a measurable map Ψ with values in $\mathscr{A}$, such that if one sets

$$\Xi_i = \Psi(\xi_{i-1}, \xi_i, \xi_{i+1})$$

for every site i in $\mathbb{Z}$, then the distribution of $(\Xi_i)_{i \in \mathbb{Z}}$ is π. In particular, any collections $(\Xi_i)_{i \in I}$ and $(\Xi_i)_{i \in J}$ are independent as soon as the subsets I and J of $\mathbb{Z}$ are such that $|i - j| \geqslant 3$ for every sites i in I and j in J. We call this property 2-dependence.

3.1.3 Notations

Our estimators are based on various quantities provided by the alignment of the two sequences.

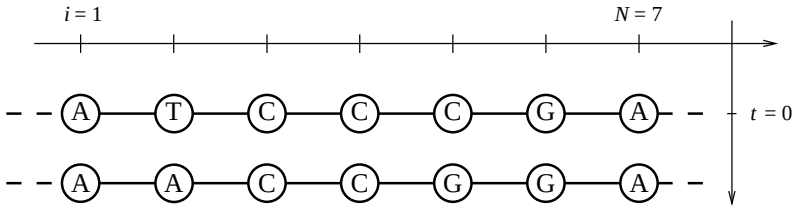


Figure 3.1: Alignment of an ancestral sequence and a present one

For every $\ell \geqslant 0$ and every word w of length $\ell + 1$ written in the alphabet $\mathscr{A}$, say that site i is occupied at time t by w if $X_{i:i+\ell}(t) = w$. For every triple of subsets W, W' and W'' of words and every couple of times t and s, $(W)(t)$ denotes the frequency of sites occupied by any of the words in W at time t, that is

$$(W)(t) = \lim_{N \to \infty} \frac{1}{N} \sum_{i=0}^{N} \sum_{w \in W} H_i(t,w), \quad \text{where} \quad H_i(t,w) = \mathbf{1}\{X_{i:i+\ell}(t) = w\},$$

and $(W, W')(t)$ the frequency of sites occupied by any of the words in W at time 0

and any of the words in W' at time t, that is

$$(W,W')(t) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N} \sum_{w\in W} \sum_{w'\in W'} H_i(0,w)H_i(t,w').$$

The limits above exist thanks to the ergodicity of π with respect to translations.

When comparing two present sequences, we use the following notations. For every sets W and W' of words and every time t, $[W,W'](t)$ denotes the frequency of sites occupied by a word of W in the left sequence (denoted by X^1) and by a word of W' in the right sequence (denoted by X^2).

We identify a word w and the set of words $\{w\}$. For every letter x in the alphabet $\mathscr{A}$, we use the shorthands $*x = \mathscr{A}\times\{x\}$, $x* = \{x\}\times\mathscr{A}$, $x*x = \{x\}\times\mathscr{A}\times\{x\}$ and $\bar{x} = \mathscr{A}\setminus\{x\}$.

3.2 Summary of main results

Our main result is theorem 3.2.4 below, which provides asymptotic confidence intervals for an estimation procedure of the time elapsed between a present sequence and an ancestral one and for the time since two present sequences issued from the same ancestral sequence diverged, for the Jukes-Cantor model with CpG influence (JC+CpG) of intensity r. These intervals are based on two consistent estimators of the elapsed time and two consistent estimators of the time of divergence.

3.2.1 Estimators and asymptotic confidence intervals

Our first estimator is based on the evolution of the frequency $(C,C)(t)$ when the time t varies and the second one on the evolution of $(A,A)(t)$. These estimators match the classic ones used for the original Jukes-Cantor model when $r = 0$. The symmetry of the roles played by A and T, or by C and G in the JC+CpG model immediately gives the relations $(A,A)(t) = (T,T)(t)$ and $(G,G)(t) = (C,C)(t)$.

Our estimators for the divergence time are based on the evolution of the frequency $[C,C](t)$ when the time t varies and on the evolution of $[A,A](t)$. Even if the results are given in the same theorem, there is a substantial difference between $[C,C]$ and $[A,A]$. Indeed, as we explain in sections 3.5 and 3.6:

Theorem 3.2.1. *In the JC+CpG model, for every positive t,*

$$[C,C](t) = (C,C)(2t), \quad [A,A](t) \neq (A,A)(2t).$$

In section 3.8, theorem 3.8.1 provides an asymptotic confidence interval for our estimation procedure of the time elapsed between a present sequence and an ancestral

one, for RN+YpR models, under the condition that the estimator is well-defined in the general case.

The keystep for the creation of phylogenetic trees built by a distance-based method is theorem 3.2.4 below. At the moment, a prior knowledge of the parameter r is needed to apply the method. In theory, one could estimate r as well, using sufficiently many quantities $(w,w')(t)$ or $[w,w'](t)$ for (small) words w and w' (in practice, it should be enough to consider nucleotides and dinucleotides w and w'). For instance, in JC+CpG, modulo some monotonicity properties, two "independent" functions $(w,w')(t)$ or $[w,w'](t)$ should be enough to estimate t and r at the same time. A different approach to estimate all the parameters at the same time, based on maximum likelihood principle and on the independence properties of the RN+YpR models at stationarity, is currently developed by Bérard and Guéguen [BG10] in the context of the alignment of genomic sequences of Human, Chimpanzee, and Macaque.

We now introduce some notations needed to state theorem 3.2.4 and used in the rest of the chapter.

Definition 3.2.2. *Let $(x,x)^N_{\text{obs}}$ and $[x,x]^N_{\text{obs}}$ denote for every $x \in \{A,C\}$ the observed value of (x,x) and $[x,x]$ on two aligned sequences of length N, that is,*

$$(x,x)^N_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N}(K_i^x)(t), \quad \textit{with} \quad (K_i^x)(t) = \mathbf{1}\{X_i(0) = X_i(t) = x\},$$

and

$$[x,x]^N_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N}[K_i^x](t), \quad \textit{with} \quad [K_i^x](t) = \mathbf{1}\{X_i^1(t) = X_i^2(t) = x\}.$$

In figure 3.1 for instance, $N = 7$ and $(C,C)^N_{\text{obs}} = \frac{2}{7}$. The quantity $(x,x)^N_{\text{obs}}$ plays the same role in the JC+CpG model than P_{obs}, introduced in chapter 2, in the classical Jukes-Cantor model.

The quantity $(x,x)^N_{\text{obs}}$ is theoretical, because generally, we do not have the ancestral DNA sequence. However, like we do in the Jukes-Cantor model, it is an intermediate tool to study $[x,x]^N_{\text{obs}}$ and provide an estimator of the divergence time between two homologous DNA sequences.

Definition 3.2.3. *Let (T_x^N) and $[T_x^N]$ denote the estimators of the elapsed time and the divergence time respectively, defined for every $x \in \{A,C\}$, as the solution in t of the equations*

$$(x,x)(t) = (x,x)^N_{\text{obs}} \quad \textit{and} \quad [x,x](t) = [x,x]^N_{\text{obs}}.$$

For $x \in \{A,C\}$, let $(\kappa_x^N)_{\text{obs}}$, $[\kappa_x^N]_{\text{obs}}$, $(\nu_x^N)_{\text{obs}}$ and $[\nu_x^N]_{\text{obs}}$ denote observed quantities,

defined as

$$
\begin{aligned}
(\kappa_C^N)_{\text{obs}} &= 4(C,C)_{\text{obs}}^N + r(C*,CG)_{\text{obs}}^N - (C)_*,\\
(\kappa_A^N)_{\text{obs}} &= 4(A,A)_{\text{obs}}^N - r(*A,CG)_{\text{obs}}^N - (A)_*,\\
(\nu_x^N)_{\text{obs}} &= (x,x)_{\text{obs}}^N - 5(x,x)_{\text{obs}}^{N\ 2} + 2(xx,xx)_{\text{obs}}^N + 2(x*x,x*x)_{\text{obs}}^N,
\end{aligned}
$$

and

$$
\begin{aligned}
[\kappa_C^N]_{\text{obs}} &= 8[C,C]_{\text{obs}}^N + 2r[C*,CG]_{\text{obs}}^N - 2(C)_*,\\
[\kappa_A^N]_{\text{obs}} &= 8[A,A]_{\text{obs}}^N - 2r[*A,CG]_{\text{obs}}^N - 2(A)_*,\\
[\nu_x^N]_{\text{obs}} &= [x,x]_{\text{obs}}^N - 5[x,x]_{\text{obs}}^{N\ 2} + 2[xx,xx]_{\text{obs}}^N + 2[x*x,x*x]_{\text{obs}}^N,
\end{aligned}
$$

where $(x)_*$ denotes the frequency of x at stationarity in the JC+CpG model.

The quantities $(\kappa_x^N)_{\text{obs}}$ play the role of $3-4P_{\text{obs}}$ in the classical Jukes-Cantor model. Indeed, we prove in lemma 3.4.1, that $(\kappa_x^N)_{\text{obs}}$ converges almost surely to $-(x,x)'(t)$. Similarly, the quantities $(\nu_x^N)_{\text{obs}}$ play the role of $\sqrt{P_{\text{obs}}(1-P_{\text{obs}})}$ in the classical Jukes-Cantor model. Finally, (T_x^N) plays the role of the quantity we denoted by D in the classical Jukes-Cantor model.

We note that $(\kappa_x^N)_{\text{obs}}$, $[\kappa_x^N]_{\text{obs}}$, $(\nu_x^N)_{\text{obs}}$ and $[\nu_x^N]_{\text{obs}}$ may be negative for some sequences of observations and some lengths N. However, from lemma 3.4.1, $(\kappa_x^N)_{\text{obs}}$, $[\kappa_x^N]_{\text{obs}}$, $(\nu_x^N)_{\text{obs}}$ and $[\nu_x^N]_{\text{obs}}$ are almost surely positive when N is large.

As explained in sections 3.5 and 3.6, in JC+CpG, for every x, the functions

$$t \mapsto (x,x)(t), \quad \text{and} \quad t \mapsto [x,x](t),$$

are decreasing functions of $t \geqslant 0$, from $(x)_*$ at $t=0$ to $(x)_*^2$ at $t=+\infty$. Thus, (T_x^N) and $[T_x^N]$ are unique and well defined for any pair of aligned sequences such that

$$(x)_*^2 < (x,x)_{\text{obs}}^N, [x,x]_{\text{obs}}^N < (x)_*.$$

Thanks to the ergodicity of the model, this condition is almost surely satisfied when N is large enough because $(x,x)_{\text{obs}}^N \to (x,x)(t)$ and $[x,x]_{\text{obs}}^N \to [x,x](t)$ almost surely when $N \to \infty$.

However, even if (T_x^N) and $[T_x^N]$ are unique and well defined, the formulas to compute them are not straightforward since these involve the inverse of a function. Thus, to solve equation $(x,x)(t) = (x,x)_{\text{obs}}^N$, for example, one has to rely on numerical methods. Fortunately, explicit formulas for $(x,x)(t)$ and $[x,x](t)$ in JC+CpG do exist.

We now state our main result.

Theorem 3.2.4. *Assume that the ancestral sequence is at stationarity. Then, in JC+CpG, for $x=A$ and $x=C$,*

$$(\kappa_x^N)_{\text{obs}}\sqrt{N/(\nu_x^N)_{\text{obs}}}\left((T_x^N)-t\right) \quad \textit{and} \quad [\kappa_C^N]_{\text{obs}}\sqrt{N/[\nu_C^N]_{\text{obs}}}\left([T_C^N]-t\right)$$

both converge in distribution to the standard normal law when $N \to +\infty$.

In the ancestral case, an asymptotic confidence interval at level ε for the elapsed time is

$$\left[(T_x^N) - \frac{z(\varepsilon)}{(\kappa_x^N)_{\text{obs}}}\sqrt{\frac{(v_x^N)_{\text{obs}}}{N}}, (T_x^N) + \frac{z(\varepsilon)}{(\kappa_x^N)_{\text{obs}}}\sqrt{\frac{(v_x^N)_{\text{obs}}}{N}}\right].$$

In the homologous case, an asymptotic confidence interval at level ε for the time of divergence is

$$\left[[T_C^N] - \frac{z(\varepsilon)}{[\kappa_C^N]_{\text{obs}}}\sqrt{\frac{[v_C^N]_{\text{obs}}}{N}}, [T_C^N] + \frac{z(\varepsilon)}{[\kappa_C^N]_{\text{obs}}}\sqrt{\frac{[v_C^N]_{\text{obs}}}{N}}\right].$$

In both formulas, $z(\varepsilon)$ denotes the unique real number such that $\mathbb{P}(|Z| \geqslant z(\varepsilon)) = \varepsilon$ with Z a standard normal random variable.

Remark 3.2.5. *Note that if conjecture 3.6.6 holds, that is, the function $t \mapsto [A,A](t)$ is a decreasing diffeomorphism, then $[\kappa_A^N]_{\text{obs}}\sqrt{N/[v_A^N]_{\text{obs}}}\left([T_A^N] - t\right)$ converges in distribution to the standard normal law.*

Remark 3.2.6. *Theorem 3.2.4 implies that, for large N, the width of the confidence interval scales as $N^{-1/2}$ times a function of t, and that, for large t, this function of t scales as e^{4t} in the ancestral case and as e^{8t} in the homologous case (according to formulas given in corollaries 3.5.1 and 3.6.2). Heuristically, this means that, to estimate the time t up to a given factor, one must observe a part of the sequence of length N of order at least e^{8t} in the ancestral case and at least e^{16t} in the homologous case.*

The rest of the chapter is organized as follows. In section 3.3, we state central limit theorems for the time estimators for JC+CpG and for the general model under conjecture 3.3.4. In section 3.4, we show that the central limit theorems established in section 3.3 imply theorem 3.2.4 of section 3.2. In section 3.5, and 3.6, we characterize the evolutions of $(x,x)(t)$ and $[x,x](t)$ for $x = C$ and $x = A$, and we state some monotonicity properties.

In section 3.7, we give a short description of the general RN model with YpR influence (RN+YpR). In section 3.8, we give an extension of theorem 3.2.4 to the general model under conjecture 3.3.4, and in section 3.9 the justification of this extension. In section 3.10, we describe some simulations supporting conjecture 3.3.4.

3.3 Central limit theorems for time estimators

We give here central limit theorems for the time estimators in the general model. The strategy is the following. We first deal with $(x,x)^N_{\text{obs}}$ and $[x,x]^N_{\text{obs}}$. We compute exactly the variance of these quantities thanks to the 2-dependence. Then, we use

a central limit theorem for mixing sequences. To state central limit therorem for the time estimators, we use the delta method, and to do that, we need to know that $t \mapsto (x,x)(t)$ and $t \mapsto [x,x](t)$ are diffeomorphisms. This is still a conjecture for the general model whereas we prove it for JC+CpG.

3.3.1 Variance computations

We detail the properties of $(C,C)^N_{\text{obs}}$, $(A,A)^N_{\text{obs}}$, $[C,C]^N_{\text{obs}}$ and $[A,A]^N_{\text{obs}}$. We assume that $N \geqslant 2$.

Lemma 3.3.1. *Assume that the ancestral sequence is at stationarity. In RN+YpR, for $x \in \{C,A\}$, the mean of $(x,x)^N_{\text{obs}}$, respectively $[x,x]^N_{\text{obs}}$, with respect to π is $(x,x)(t)$, respectively $[x,x](t)$.*

The variances of $(x,x)^N_{\text{obs}}$ and $[x,x]^N_{\text{obs}}$ with respect to π are equal to $(\sigma^2_x)(N,t)$ and $[\sigma^2_x](N,t)$, where

$$\begin{aligned} N(\sigma^2_x)(N,t) =& (x,x)(t) - (x,x)(t)^2 + 2(1-1/N)\big((xx,xx)(t) - (x,x)(t)^2\big) + \\ &+ 2(1-2/N)\big((x*x,x*x)(t) - (x,x)(t)^2\big), \end{aligned}$$

and,

$$\begin{aligned} N[\sigma^2_x](N,t) =& [x,x](t) - [x,x](t)^2 + 2(1-1/N)\big([xx,xx](t) - [x,x](t)^2\big) + \\ &+ 2(1-2/N)\big([x*x,x*x](t) - [x,x](t)^2\big), \end{aligned}$$

Proof. Since the initial sequence $X(0) = (X_i(0))_{i\in\mathbb{Z}}$ is at stationarity, that is, distributed along π defined in section 3.1, then, for every positive t, $X(t) = (X_i(t))_{i\in\mathbb{Z}}$ is distributed along π.

As a consequence, the random variables $((K^x_i)(t))_{i\in\mathbb{Z}}$, respectively $([K^x_i](t))_{i\in\mathbb{Z}}$, are Bernoulli random variables identically distributed with respect to π. Their common mean is $(x,x)(t)$, respectively $[x,x](t)$.

On the other hand, $(x,x)^N_{\text{obs}}$, respectively $[x,x]^N_{\text{obs}}$, is the empirical mean of the N values $(K^x_i)(t)$, respectively $[K^x_i](t)$, for i from 1 to N. Thus, we obtain the value of $\mathbb{E}((x,x)^N_{\text{obs}})$, respectively $\mathbb{E}([x,x]^N_{\text{obs}})$, as $(x,x)(t)$, respectively $[x,x](t)$.

Furthermore,

$$N^2(\sigma^2_x)(N,t) = \sum_{i=1}^{N} \text{var}((K^x_i)(t)) + 2 \sum_{1\leqslant i<j\leqslant N} \text{cov}((K^x_i)(t),(K^x_j)(t)).$$

The variance of each $(K^x_i)(t)$ is $\text{var}((K^x_1)(t)) = (x,x)(t) - (x,x)(t)^2$.

The 2-dependence, valid for RN+YpR, implies that each covariance for $|i-j| \geqslant 3$ is zero. The invariance by translation of π, valid for RN+YpR, shows that each of the $(N-1)$ covariances such that $i = j-1$ is

$$\text{cov}((K^x_1)(t),(K^x_2)(t)) = (xx,xx)(t) - (x,x)(t)^2.$$

Finally, each of the $(N-2)$ covariances such that $i = j-2$ is

$$\mathrm{cov}((K_1^x)(t),(K_3^x)(t)) = (x*x,x*x)(t) - (x,x)(t)^2.$$

The same arguments hold for the variance of $[x,x]^N_{\mathrm{obs}}$. This concludes the proof. □

3.3.2 Central limit theorems for $(x,x)^N_{\mathrm{obs}}$ and $[x,x]^N_{\mathrm{obs}}$

We prove the convergence in distribution to the normal law, using a classical strategy based on the following result.

Theorem 3.3.2 (Hall and Heyde [HH80])**.** *Let $(V_i)_{i\in\mathbb{Z}}$ denote a stationary, ergodic, centered, square integrable sequence. Let $\mathscr{F}_0 = \sigma(V_i\,;\, i \leqslant 0)$ denote the σ-algebra generated by the random variables V_i for $i \leqslant 0$. For every positive integer n, introduce*

$$U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} V_i.$$

Assume that

(i) *for every positive n, the series $\sum_{k\geqslant 1} \mathbb{E}(V_k \mathbb{E}(V_n|\mathscr{F}_0))$ converges,*

(ii) *the series $\sum_{k\geqslant K} |\mathbb{E}(V_k \mathbb{E}(V_n|\mathscr{F}_0))|$ converges to zero when $n \to +\infty$, uniformly with respect to K.*

Then $\mathbb{E}(U_n^2)$ converges to a real number $\sigma^2 \geqslant 0$ when $n \to +\infty$. Furthermore, if $\sigma^2 > 0$, then $U_n/\sqrt{\sigma^2}$ converges in distribution to the standard normal distribution.

Proposition 3.3.3. *In RN+YpR, for $x \in \{C,A\}$, when $N \to +\infty$,*

$$\sqrt{N}((x,x)^N_{\mathrm{obs}} - (x,x)(t)) \quad \textit{and} \quad \sqrt{N}([x,x]^N_{\mathrm{obs}} - [x,x](t))$$

both converge in distribution to the centered normal distribution with variance $(\sigma_x^2)(t)$, respectively $[\sigma_x^2](t)$, where

$$(\sigma_x^2)(t) = (x,x)(t) + 2(xx,xx)(t) + 2(x*x,x*x)(t) - 5(x,x)(t)^2,$$
$$[\sigma_x^2](t) = [x,x](t) + 2[xx,xx](t) + 2[x*x,x*x](t) - 5[x,x](t)^2.$$

Proof. For any RN+YpR model, for $x \in \{C,A\}$, the sequence $((K_i^x)(t))_{i\in\mathbb{Z}}$, respectively $([K_i^x](t))_{i\in\mathbb{Z}}$, is stationary and ergodic. Introduce

$$(V_i^x) = (K_i^x)(t) - (x,x)(t), \qquad [V_i^x] = [K_i^x](t) - [x,x](t).$$

This defines a sequence $((V_i^x))_{i\in\mathbb{Z}}$, respectively $([V_i^x])_{i\in\mathbb{Z}}$, such that the first hypothesis of theorem 3.3.2 holds. We now check conditions (i) et (ii).

The 2-dependence, valid for any RN+YpR model, implies that, for every $n \geqslant 3$, $\mathbb{E}((V_n^x)|\mathscr{F}_0^x) = \mathbb{E}((V_n^x)) = 0$, respectively $\mathbb{E}([V_n^x]|\mathscr{F}_0^x) = \mathbb{E}([V_n^x]) = 0$. Hence we only have to check the cases $n = 1$ and $n = 2$.

For every $k \geqslant 3$, (V_k^x), respectively $[V_k^x]$, is independent of $\mathscr{F}_0^x$, and $\mathbb{E}((V_n^x)|\mathscr{F}_0^x)$, respectively $\mathbb{E}([V_n^x]|\mathscr{F}_0^x)$, is $\mathscr{F}_0^x$-measurable, hence

$$\mathbb{E}((V_k^x)\mathbb{E}((V_n^x)|\mathscr{F}_0^x)) = \mathbb{E}((V_k^x))\mathbb{E}(\mathbb{E}(V_n^x|\mathscr{F}_0^x)) = 0,$$

and

$$\mathbb{E}([V_k^x]\mathbb{E}([V_n^x]|\mathscr{F}_0^x)) = \mathbb{E}([V_k^x])\mathbb{E}(\mathbb{E}([V_n^x]|\mathscr{F}_0^x)) = 0.$$

This implies (i) and (ii), hence theorem 3.3.2 applies.

To compute the asymptotic variance in the theorem, we note that the variances of

$$\sqrt{N}((x,x)_{\text{obs}}^N - (x,x)(t)) \quad \text{and} \quad \sqrt{N}([x,x]_{\text{obs}}^N - [x,x](t))$$

are $N(\sigma_x^2)(N,t)$ and $N[\sigma_x^2](N,t)$ respectively, which, when $N \to +\infty$, converge to $(\sigma_x^2)(t)$ and $[\sigma_x^2](t)$ respectively. □

3.3.3 Central limit theorems for (T_x^N) and $[T_x^N]$

We describe explicitly the behaviour of $(T_x^N) - t$ and $[T_x^N] - t$. To state our result, we use the central limit theorems given in proposition 3.3.3, but we now need to treat separately JC+CpG and RN+YpR.

For $x \in \{C,A\}$, let (μ_x), respectively $[\mu_x]$, denote the inverse function of $t \mapsto (x,x)(t)$, respectively $t \mapsto [x,x](t)$. That is,

$$t = (\mu_x)((x,x)(t)) = [\mu_x]([x,x](t)),$$

and (μ_x) and $[\mu_x]$ are both defined on the interval $((x)_*^2, (x)_*]$.

From propositions 3.5.4, 3.6.3 and 3.6.6, the functions $t \mapsto (x,x)(t)$ and $t \mapsto [x,x](t)$ are diffeomorphisms in JC+CpG. In RN+YpR, this is only a conjecture, supported by simulations described in section 3.10, which seem to show that the function $t \mapsto (C,C)(t)$ is indeed decreasing.

Conjecture 3.3.4. *In RN+YpR, for $x \in \{C,A\}$, the functions $t \mapsto (x,x)(t)$ and $t \mapsto [x,x](t)$ are diffeomorphisms from $[0,+\infty)$ to $((x)_*^2, (x)_*]$.*

Then,

$$(T_x^N) = (\mu_x)((x,x)_{\text{obs}}^N) \quad \text{and} \quad t = (\mu_x)((x,x)(t)),$$

and

$$[T_x^N] = [\mu_x]([x,x]_{\text{obs}}^N) \quad \text{and} \quad t = [\mu_x]([x,x](t)).$$

Besides, the derivatives of (μ_x) and $[\mu_x]$, with respect to t are

$$(\mu_x)'((x,x)(t)) = \frac{1}{(x,x)'(t)} \quad \text{and} \quad [\mu_x]'([x,x](t)) = \frac{1}{[x,x]'(t)}.$$

Using the delta method, see [vdV98], one gets the following result.

Proposition 3.3.5. *In JC+CpG, for $x \in \{C,A\}$, when $N \to +\infty$, $\sqrt{N}((T_x^N) - t)$, respectively $\sqrt{N}([T_x^N] - t)$, converges in distribution to the centered normal distribution with variance $(\sigma_x^2)(t)/(x,x)'(t)^2$, respectively $[\sigma_x^2](t)/[x,x]'(t)^2$.*

Under conjecture 3.3.4, the same results hold for the full class RN+YpR.

3.4 Proofs for JC + CpG

Proposition 3.3.5 yields the variation of (T_x^N) and $[T_x^N]$ around t for $x \in \{C,A\}$. A priori, to build a confidence interval for t from this proposition requires to know the value of $(x,x)'(t)$, respectively $[x,x]'(t)$, and of $(\sigma_x^2)(t)$, respectively $[\sigma_x^2](t)$, which all depend on the quantity t to be estimated.

As is customary, Slutsky's lemma (see [vdV98]) allows to bypass this difficulty through the observed quantities, defined in section 3.2, $(\kappa_x^N)_{\text{obs}}$ and $(v_x^N)_{\text{obs}}$, respectively $[\kappa_x^N]_{\text{obs}}$ and $[v_x^N]_{\text{obs}}$. Indeed, Slutsky's lemma states that if two sequences of random variables $(X_N)_N$ and $(Y_N)_N$ are such that $(X_N)_N$ converges in distribution to a random variable X and $(Y_N)_N$ converges in probability to a constant c, then the sequence $(X_N Y_N)_N$ converges in distribution to the random variable cX.

Lemma 3.4.1. *In JC+CpG, for $x \in \{C,A\}$,*

$$(\kappa_x^N)_{\text{obs}} \to -(x,x)'(t), \quad [\kappa_x^N]_{\text{obs}} \to -[x,x]'(t),$$

and

$$(v_x^N)_{\text{obs}} \to (\sigma_x^2)(t), \quad [v_x^N]_{\text{obs}} \to [\sigma_x^2](t)$$

almost surely when $N \to +\infty$.

Proof. The equalities

$$(C,C)'(t) = -4(C,C)(t) - r(C*,CG)(t) + (C)_*,$$
$$(A,A)'(t) = -4(A,A)(t) + r(*A,CG)(t) + (A)_*,$$

given in sections 3.5 and 3.6, and the almost sure convergence of the observed quantities

$$(C,C)^N_{\text{obs}}, \quad (C*,CG)^N_{\text{obs}}, \quad (CC,CC)^N_{\text{obs}}, \quad (C*C,C*C)^N_{\text{obs}},$$

and,

$$(A,A)^N_{\text{obs}}, \quad (*A,CG)^N_{\text{obs}}, \quad (AA,AA)^N_{\text{obs}}, (A*A,A*A)^N_{\text{obs}},$$

to the corresponding limiting values, when $N \to +\infty$, imply the desired convergences. Likewise, the equalities

$$[C,C]'(t) = -8[C,C](t) - 2r[C*,CG](t) + 2(C)_*,$$
$$[A,A]'(t) = -8[A,A](t) + 2r[*A,CG](t) + 2(A)_*,$$

imply the convergence of $[\kappa_x^N]_{\text{obs}}$. □

We apply Slutsky's lemma to $(X_N) = (\sqrt{N}((T_x^N) - t))$ and $(X_N) = (\sqrt{N}([T_x^N] - t))$, which, from proposition 3.3.5, converge in distribution to the centered normal law with variance $(\sigma_x^2)(t)/(x,x)'(t)^2$ and $[\sigma_x^2](t)/[x,x]'(t)^2$ respectively, and to $(Y_N) = ((\kappa_x^N)_{\text{obs}}/\sqrt{(v_x^N)_{\text{obs}}})$ and $(Y_N) = ([\kappa_x^N]_{\text{obs}}/\sqrt{[v^x]_{\text{obs}}})$, which converge in probability to $-(x,x)'(t)/(\sigma_x)(t)$ and $-[x,x]'(t)/[\sigma_x](t)$ respectively, from lemma 3.4.1. This implies theorem 3.2.4.

3.5 Evolutions of $(C,C)(t)$ and $[C,C](t)$ in JC+CpG

3.5.1 Dynamics of $(C,C)(t)$

In JC+CpG, the dinucleotides encoded as $\{C,\bar{C}\} \times \{G,\bar{G}\}$ have autonomous evolution with the following 4×4 rate matrix Q:

$$\begin{array}{c} \\ CG \\ \bar{C}G \\ \bar{C}\bar{G} \\ C\bar{G} \end{array} \begin{array}{c} \begin{array}{cccc} CG & \bar{C}G & \bar{C}\bar{G} & C\bar{G} \end{array} \\ \begin{pmatrix} -(6+2r) & 3+r & 0 & 3+r \\ 1 & -4 & 3 & 0 \\ 0 & 1 & -2 & 1 \\ 1 & 0 & 3 & -4 \end{pmatrix} \end{array}.$$

The dynamics of the dinucleotides can be represented with the graph given in figure 3.2.

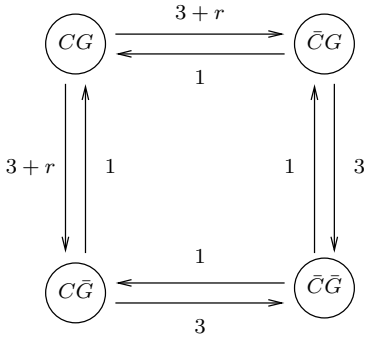


Figure 3.2: Dynamics of dinucleotides encoded as $\{C,\bar{C}\} \times \{G,\bar{G}\}$

The exponential of the corresponding matrix can be explicitly computed. Indeed, the eigenvalues of Q are respectively 0, -4, $-u_-$ and $-u_+$, with

$$u = \sqrt{4+2r+r^2}, \quad u_+ = 6+r+u, \qquad u_- = 6+r-u,$$

and the corresponding eigenvectors are

$$\begin{pmatrix}1\\1\\1\\1\end{pmatrix}, \begin{pmatrix}0\\-1\\0\\1\end{pmatrix}, \begin{pmatrix}3r+3u-ru_-\\2+r\\(2-u_+)/3\\2+r\end{pmatrix}, \begin{pmatrix}3r-3u-ru_+\\2+r\\(2-u_-)/3\\2+r\end{pmatrix}.$$

Then, an explicit expression can be obtained for every $(XY,ZT)(t)$ where (X,Y) and (Z,T) belong to $\{C,\bar{C}\}\times\{G,\bar{G}\}$. For instance,

$$\begin{aligned}
(CG,CG)(t) &= (CG)(0)\left(\alpha_0+\alpha_+\mathrm{e}^{-u_+t}+\alpha_-\mathrm{e}^{-u_-t}\right),\\
(CG,C\bar{G})(t) &= (CG)(0)\left(\beta_0+\beta_+\mathrm{e}^{-u_+t}+\beta_-\mathrm{e}^{-u_-t}\right),\\
(C\bar{G},CG)(t) &= (C\bar{G})(0)\left(\gamma_0+\gamma_+\mathrm{e}^{-u_+t}+\gamma_-\mathrm{e}^{-u_-t}\right),\\
(C\bar{G},C\bar{G})(t) &= (C\bar{G})(0)\left(\delta_0+\delta_1\mathrm{e}^{-4t}+\delta_+\mathrm{e}^{-u_+t}+\delta_-\mathrm{e}^{-u_-t}\right),
\end{aligned}$$

with

$$\alpha_0=\gamma_0=\frac{1}{16+5r},\quad \beta_0=\delta_0=\frac{3+r}{16+5r},\quad \delta_1=\frac{1}{2},$$

$$\begin{aligned}
\alpha_\pm &= \frac{1}{2u(16+5r)}\left(5u(3+r)\pm(6+17r+5r^2)\right),\\
\beta_\pm &= \frac{1}{2u(16+5r)}\left(-u(3+r)\mp(30+22r+4r^2)\right),\\
\gamma_\pm &= \frac{1}{2u(16+5r)}\left(-u\mp(10+4r)\right),\\
\delta_\pm &= \frac{1}{4u(16+5r)}\left(u(10+3r)\mp(4+8r+3r^2)\right).
\end{aligned}$$

These expressions are also valid out of equilibrium, that is, when the distribution of $X(0)$ may be different of π.

One can also compute explicitly the stationary frequencies of the dinucleotides encoded in $\{C,\bar{C}\}\times\{G,\bar{G}\}$ using the same matrix. That is

$$\begin{aligned}
(CG)_* &= \frac{1}{16+5r}, & (C\bar{G})_* &= \frac{3+r}{16+5r},\\
(\bar{C}\bar{G})_* &= \frac{9+3r}{16+5r}, & (\bar{C}G)_* &= \frac{3+r}{16+5r}.
\end{aligned}$$

These stationary frequencies are in [BGP08].

We observe that $(C,C)(t)$ can be expressed as a linear combination of terms of the form $(XY,ZT)(t)$ where (X,Y) and (Z,T) belong to $\{C,\bar{C}\}\times\{G,\bar{G}\}$. Indeed,

$$(C,C)(t)=(CG,CG)(t)+(CG,C\bar{G})(t)+(C\bar{G},CG)(t)+(C\bar{G},C\bar{G})(t).$$

Thus, an explicit expression for $(C,C)(t)$ can be obtained, for instance in the stationary regime where the initial values are $(C\bar{G})(0)=(C\bar{G})_*$ and $(CG)(0)=(CG)_*$.

Proposition 3.5.1. *In the stationary regime,*

$$(C,C)(t) = c_1 e^{-4t} + c_+ e^{-u_+ t} + c_- e^{-u_- t} + (C)_*^2,$$

with

$$c_1 = \frac{3+r}{2(16+5r)} \quad \textit{and} \quad c_\pm = \frac{3+r}{4u(16+5r)^2}\left(u(16+3r) \mp (32+14r+3r^2)\right).$$

As expected,

$$c_+ + c_- + c_1 = (C)_* - (C)_*^2.$$

Furthermore, for every positive r,

$$4 < u_- < 5 < 2r+7 < u_+ < 2r+8.$$

The expression of $(C,C)(t)$ as a linear combination of terms of the form $(XY,ZT)(t)$ where (X,Y) and (Z,T) belong to $\{C,\bar{C}\}\times\{G,\bar{G}\}$, yields an expression of $(C,C)'(t)$ in terms of dinucleotide frequencies by using the transposed matrix of Q.

Proposition 3.5.2. *The evolution of $(C,C)(t)$ satisfies the linear differential equation*

$$(C,C)'(t) = -4(C,C)(t) - r(C*,CG)(t) + (C)(0).$$

Proposition 3.5.2 is valid out of equilibrium. This proposition is necessary to prove the almost sure convergence of $(\kappa_C^N)_{\text{obs}}$ to $-(C,C)'(t)$ in lemma 3.4.1.

We now compare the dynamics of $(C,C)(t)$ in the standard Jukes-Cantor model to the dynamics in JC+CpG with the same overall rate of substitutions.

Proposition 3.5.3. *The convergence of $(C,C)(t)$ to equilibrium when $t \to +\infty$ in JC+CpG is slower than in the independent Jukes-Cantor model with the same global rate of substitution.*

We prove proposition 3.5.3 at the end of section 3.6. Now we study the dynamics of $[C,C](t)$.

3.5.2 Dynamics of $[C,C](t)$

Although JC+CpG is not reversible, the dynamics of dinucleotides encoded as $\{C,\bar{C}\}\times\{G,\bar{G}\}$ is reversible. Indeed, on figure 3.2, one can see that there is only one cycle in the graph, and that the product of the rates is the same whatever direction in the loop is chosen. Hence, Kolmogorov criterium holds and the dynamics is reversible.

Reversibility means that the dynamics will look the same whether time runs forward or backward. As a result, given two sequences at stationarity (without stationarity, this is wrong), the probability of data in a state is the same whether

one sequence is ancestral to the other or both are descendants of an ancestral sequence at stationarity. Roughly speaking, for every (X,Y) and (Z,T) that belong to $\{C,\bar{C}\}\times\{G,\bar{G}\}$, going from a XY at time t to 0 then back to a ZT at time t on another branch, is equivalent to going from a XY to at time 0 to a ZT at time $2t$.

As a consequence, for every positive t, we have

$$[C,C](t) = (C,C)(2t).$$

The equality above is not satisfied for every word of the sequence. For instance, in section 3.6, we prove that, as soon as $r > 0$ and $t > 0$,

$$[A,A](t) \neq (A,A)(2t).$$

If JC+CpG was reversible, the equality above would be true, but we insist on the fact that JC+CpG is not reversible, and that the "miracle" equality $[C,C](t) = (C,C)(2t)$ is a consequence of the reversibility of the dynamics of dinucleotides encoded as $\{C,\bar{C}\}\times\{G,\bar{G}\}$.

For every positive r, the parameters $c_{\pm}$ and c_1, are positive. This proves the following proposition.

Proposition 3.5.4. *In JC+CpG, the functions $t \mapsto (C,C)(t)$ and $t \mapsto [C,C](t)$ are decreasing diffeomorphisms from $[0,+\infty)$ to $((C)^2_*,(C)_*]$.*

3.6 Evolutions of $(A,A)(t)$ and $[A,A](t)$ in JC+CpG

Like we did to study (C,C), it is possible to encode dinucleotides such that in JC+CpG, (A,A) is a linear combination of terms involved in an autonomous evolution. It suffices to encode the dinucleotides as $\{C,\bar{C}\}\times\{A,G,Y\}$, and the dynamics can be represented with the graph given in figure 3.3.

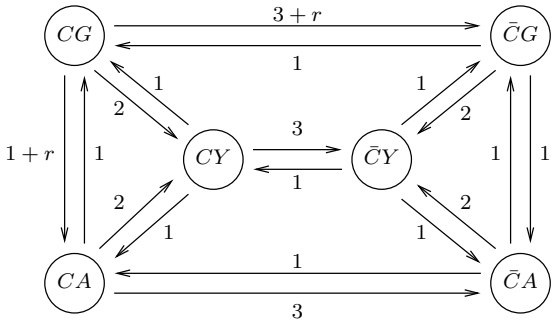


Figure 3.3: Dynamics of dinucleotides encoded as $\{C,\bar{C}\}\times\{A,G,Y\}$

However, we will not use this encoding to compute $(A,A)(t)$. Indeed, the evolution matrix associated to this encoding is a 6×6 matrix whereas it is possible to deal

with the 4×4 matrix Q, defined in section 3.5, to characterize the evolution of (A,A). We now explain this.

We chose to present this encoding because it is a way to understand the difference between the role of C and A in the Jukes Cantor model with CpG effect. Indeed, the dynamics of dinucleotides encoded as $\{C,\bar{C}\} \times \{A,G,Y\}$ is not reversible. This can be checked by looking at the cycle $CA \to CY \to CG \to CA$ in figure 3.3. As a consequence, even if the non-reversibility of the dynamics does not strictly prove that the identity $[A,A](t) = (A,A)(2t)$ never holds when $r > 0$, the non reversibility of the dynamics can explain why such an identity is unlikely to be true, and in fact, unlike in the case of $[C,C]$, as soon as $r > 0$ and $t > 0$,

$$[A,A](t) \neq (A,A)(2t).$$

We explain this at the end of the current section. Now, we describe a way to state the expression of $(A,A)(t)$.

3.6.1 Dynamics of $(A,A)(t)$

Proposition 3.6.1. *The evolution of $(A,A)(t)$ satisfies the linear differential equation*

$$(A,A)'(t) = -4(A,A)(t) + r(*A,CG)(t) + (A)(0).$$

Proposition 3.6.1 is necessary to prove the almost sure convergence of $(\kappa_A^N)_{\text{obs}}$ to $-(A,A)'(t)$.

Proof of proposition 3.6.1. To compute $(A,A)'(t)$, one must compare $(A,A)(t+s)$ to $(A,A)(t)$ up to the order s, for every positive t and vanishingly small positive s. The probability that at least two substitutions occur at the same site between times t and $t+s$ is $o(s)$, hence these events do not appear in the limit we consider.

Since in a dinucleotide CpG, only nucleotide G can lead to nucleotide A with additional rate, we only consider the set of two-letter configurations leading to different transition rates to $*A$. For every positive t and s,

$$(A,A)(t+s) = (*A,CA)(t+s) + (*A,\bar{C}A)(t+s).$$

On the other hand, let $(W'|W)(s)$ denote the probability that sites occupied by a word of W at time 0 are occupied by a word of W' at time s. Then,

$$\begin{aligned}
&(CA|CY)(s) = s + o(s), \quad (CA|CA)(s) = 1 - 6s + o(s),\\
&(CA|\bar{C}A)(s) = s + o(s), \quad (CA|CG)(s) = (1+r)s + o(s),\\
&(\bar{C}A|\bar{C}\bar{A})(s) = s + o(s), \quad (\bar{C}A|CA)(s) = 3s + o(s),
\end{aligned}$$

$$(\bar{C}A|\bar{C}A)(s) = 1 - 4s + o(s).$$

We are now ready to evaluate $(A,A)(t+s)$. Decomposing along the values at time t, and using the Markov property, one gets

$$(*A,CA)(t+s) = (*A,CY)(t)s + (*A,CG)(t)(1+r)s + (*A,CA)(t)(1-6s) + \\ + (*A,\bar{C}A)(t)s + o(s),$$

and,

$$(*A,\bar{C}A)(t+s) = (*A,\bar{C}\bar{A})(t)s + (*A,\bar{C}A)(t)(1-4s) + (*A,CA)(t)3s + o(s).$$

Using the relations

$$\begin{aligned}(*A,CY)(t) + (*A,CG)(t) &= (*A,C\bar{A})(t),\\ (*A,\bar{C}\bar{A})(t) + (*A,C\bar{A})(t) &= (A,\bar{A})(t),\\ (*A,CA)(t) + (*A,\bar{C}A)(t) &= (A,A)(t),\end{aligned}$$

one gets finally,

$$(A,A)(t+s) = (A,\bar{A})(t)s + r(*A,CG)(t) + (1-3s)(A,A)(t) + o(s).$$

The sum of the contributions of order 1 is $(A,A)(t)$. The sum of the contributions of order s yields the derivative, hence

$$(A,A)'(t) = -3(A,A)(t) + (A,\bar{A})(t) + r(*A,CG)(t).$$

Using the relation

$$(A,A)(t) = (A)(0) - (A,\bar{A})(t),$$

one gets the result. □

Let $U(t)$ denote the time dependent vector defined as

$$\begin{pmatrix}(*A,CG)(t)\\ (*A,\bar{C}G)(t)\\ (*A,\bar{C}\bar{G})(t)\\ (*A,C\bar{G})(t)\end{pmatrix},$$

then we have, as a straightforward consequence of the encoding $\{C,\bar{C}\}\times\{G,\bar{G}\}$,

$$U'(t) = ({}^tQ)\cdot U(t).$$

We can now compute $(*A,CG)(t)$, infer the value $(A)_*$ of $(A)(0)$ at stationarity and finally state the expression of $(A,A)(t)$.

Corollary 3.6.2. *In the stationary regime,*

$$(A,A)(t) = a_1 e^{-4t} + a_+ e^{-u_+ t} + a_- e^{-u_- t} + (A)_*^2,$$

with

$$a_1 = \frac{80+31r}{32(16+5r)},$$

and

$$a_\pm = \frac{512+384r+106r^2+13r^3 \mp u(256+18r+13r^2)}{64u(16+5r)^2}.$$

For every positive r, the parameters $a_\pm$ and a_1, are positive. This proves the following proposition.

Proposition 3.6.3. *In JC+CpG at stationarity, the function $t \mapsto (A,A)(t)$ is a decreasing diffeomorphism from $[0,+\infty)$ to $((A)_*^2,(A)_*]$.*

We now compare the dynamics of $(A,A)(t)$ in the standard Jukes and Cantor model with the Jukes-Cantor model with CpG influence with the same overall rate of substitutions.

Proposition 3.6.4. *The convergence of $(A,A)(t)$ to equilibrium when $t \to +\infty$ in the Jukes-Cantor model with CpG influence is slower than in the independent Jukes-Cantor model with the same global rate of substitution.*

We prove proposition 3.6.4 at the end of the current section. We deal now with the evolution of $[A,A](t)$.

3.6.2 Dynamics of $[A,A](t)$

Extending the strategy used to prove corollary 3.6.2, one can also derive an explicit expression for $[A,A](t)$.

Introduce the constant matrices

$$M = \begin{pmatrix} -8 & 2r & 0 & 0 \\ 0 & -(12+2r) & 1 & 1 \\ 0 & -r & -8 & 0 \\ 0 & -r & 0 & -8 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & r & 0 \\ 0 & r & 0 & 0 \\ 0 & r & 0 & 0 \end{pmatrix},$$

$$O = \begin{pmatrix} -8 & -2r & 0 & 0 \\ 1 & -(12+2r) & -r & 1 \\ 0 & 4 & -(16+4r) & 0 \\ 0 & -2r & 0 & -8 \end{pmatrix}, \quad B = \begin{pmatrix} 2(A)(0) \\ (CG)(0) \\ (C)(0)+(A)(0) \\ (A)(0)+(G)(0) \end{pmatrix},$$

$$D = \begin{pmatrix} 2(C)(0) \\ (CG)(0) \\ 0 \\ 2(C)(0) \end{pmatrix},$$

and the time-dependent vectors

$$V(t) = \begin{pmatrix} [A,A](t) \\ [*A,CG](t) \\ [*A,C*](t) \\ [A,G](t) \end{pmatrix}, \quad W(t) = \begin{pmatrix} [C,C](t) \\ [C*,CG](t) \\ [CG,CG](t) \\ [C*,*G](t) \end{pmatrix}.$$

Proposition 3.6.5. *The evolution of $V(t)$ and $W(t)$ is ruled by the linear differential systems*

$$V'(t) = MV(t) + NW(t) + B, \quad W'(t) = OW(t) + D.$$

The exponential of the block matrix $\begin{pmatrix} M & N \\ 0 & O \end{pmatrix}$ can be explicitly computed. The explicit spectrum of this matrix is

$$\{-8,-8,-8,-v_+,-v_-,-2u_+,-2u_-,-(12+2r)\}, \quad v_\pm = 10 + r \pm u.$$

Thus, it is possible to compute an explicit expression for $[A,A](t)$. The computation under `Maple` yields

$$[A,A](t) = b_0 + b_\pm e^{-u_\pm t} + d_\pm e^{-v_\pm t},$$

with

$$b_0 = \frac{16}{(16+5r)^2},$$

$$b_\pm = \frac{1}{32(16+5r)^2}\left(u(384+200r+24r^2) \mp (768+592r+184r^2+24r^3)\right),$$
$$d_\pm = \frac{1}{32(16+5r)^2}\left(-u(512+272r+35r^2) \pm (1024+800r+262r^2+35r^3\right).$$

This computation shows that the coefficients of $e^{-v_+ t}$ and $e^{-v_- t}$ in the expression of $[A,A](t)$ are nonzero. This fact alone proves that $[A,A](t)$ cannot be equal to $(A,A)(2t)$ for every t. However, we observe on simulations that the two quantities are very close. On figure 3.4, one can see the simulation performed for $r = 10$, and the largest difference between $[A,A](t)$ and $(A,A)(t)$ is about 7.10^{-4}.

For every positive r, the parameter d_- is negative. Hence, we cannot prove as easily as for $t \mapsto (A,A)(t)$ that $t \mapsto [A,A](t)$ is a decreasing diffeomorphism. The derivative of $[A,A](t)$ is also useless to prove this. However, we perform some simulations, illustrated on figure 3.5, to support the following conjecture. We explain in chapter 5, how it may be possible to prove it.

Conjecture 3.6.6. *In JC+CpG, the function $t \mapsto [A,A](t)$ is a decreasing diffeomorphism from $[0,+\infty)$ to $((A)^2_*, (A)_*]$.*

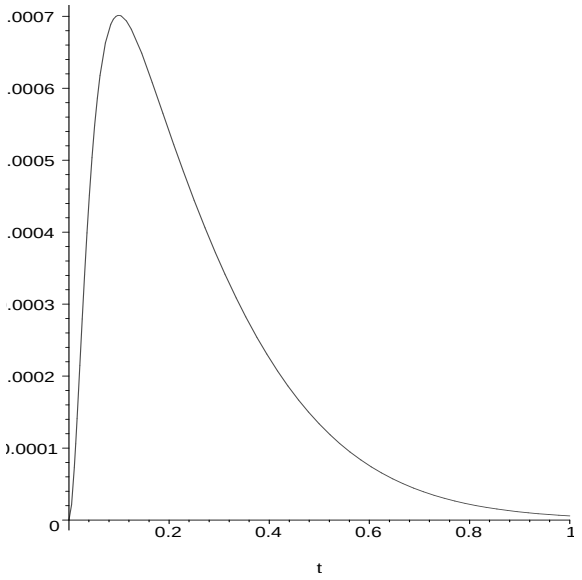


Figure 3.4: Representation of $t \mapsto [A,A](t) - (A,A)(2t)$, when $r = 10$

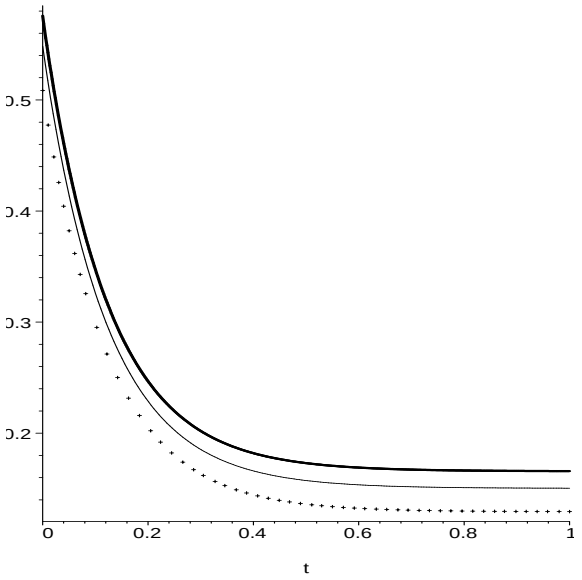


Figure 3.5: Representation of $t \mapsto [A,A](t)$, when $r = 10$ (thick line), $r = 3$ (normal line) and $r = 0.3$ (dashed line).

3.6.3 Proof of propositions 3.5.3 and 3.6.4

From corollaries 3.5.1 and 3.6.2, for every value of r, the convergence of $(C,C)(t)$ and $(A,A)(t)$ to equilibrium when $t \to +\infty$ in JC+CpG are like e^{-4t}.

In JC+CpG, every nucleotide changes at rate 3 due to unconditional substitution rates, plus every dinucleotide CpG changes at rate $2r$. Hence the global rate of substitution is

$$3+2r(CG)_* = 3+\frac{2r}{16+5r}.$$

On the other hand, in the independent Jukes-Cantor model of parameter λ, the global rate of substitution is 3λ. Hence one should set

$$\lambda = 1+\frac{2r/3}{16+5r}.$$

For independent Jukes-Cantor models, [Yan06] computes

$$(C,C)(t) = (A,A)(t) = \frac{1}{16}+\frac{3}{16}\mathrm{e}^{-4\lambda t}.$$

Since $\lambda > 1$ for every $r > 0$, the comparison with the independent Jukes-Cantor model is done.

3.7 Short description of RN+YpR and notations

First, RN stands for Rzhetsky-Nei and means that the 4×4 matrix of substitution rates which characterize the independent evolution of the sites must satisfy 4 equalities, summarized as follows: for every pair of nucleotides x and $y \neq x$, the substitution rate from x to y may depend on x but only through the fact that x is a purine (A or G, symbol R) or a pyrimidine (C or T, symbol Y). For instance, the substitution rates from C to A and from T to A must coincide, likewise for the substitution rates from A to C and from G to C, from C to G and from T to G, and finally from A to T and from G to T. The 4 remaining rates, corresponding to purine-purine substitutions and to pyrimidine-pyrimidine substitutions, are free.

Second, the influence mechanism is called YpR, which stands for the fact that one allows any specific substitution rates between any two YpR dinucleotides (CG, CA, TG and TA) which differ by one position only, for a total of 8 independent parameters. JC+CpG is the simplest non trivial case: the only YpR substitutions with positive rate are $CG \to CA$ and $CG \to TG$, and both happen at the same rate.

Recall that Y denote the set of pyrimidines defined as $Y = \{T,C\}$, and R the set of purines defined as $= \{A,G\}$.

The 4×4 matrix of substitution rates which characterize the independent evolution of the sites in RN model is given by

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array}\!\!\begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & v_T & v_C & w_G \\ v_A & \cdot & w_C & v_G \\ v_A & w_T & \cdot & v_G \\ w_A & v_T & v_C & \cdot \end{pmatrix} \end{array}.$$

The influence mechanism called YpR adds specific rates of substitutions from each YpR dinucleotide as follows.

- Every dinucleotide CG moves to CA at rate r_A^C and to TG at rate r_T^G.
- Every dinucleotide TA moves to CA at rate r_C^A and to TG at rate r_G^T.
- Every dinucleotide CA moves to CG at rate r_G^C and to TA at rate r_T^A.
- Every dinucleotide TG moves to CG at rate r_C^G and to TA at rate r_A^T.

3.8 Extension of theorem 3.2.4 to RN+YpR

Under conjecture 3.3.4, it is possible to generalize theorem 3.2.4 by suitably generalizing the definitions of κ and v given in section 3.2. Introduce the parameters

$$\begin{aligned}(\kappa_{RN}^N)_{\text{obs}} = & -v_C(C,A)_{\text{obs}}^N - w_C(C,T)_{\text{obs}}^N + (v_A + w_T + v_G)(C,C)_{\text{obs}}^N - v_C(C,G)_{\text{obs}}^N \\ & - r_C^A(C*,TA)_{\text{obs}}^N - r_C^G(C*,TG)_{\text{obs}}^N + r_T^A(C*,CA)_{\text{obs}}^N + r_T^G(C*,CG)_{\text{obs}}^N. \\ (v_{RN}^N)_{\text{obs}} = & (v_C^N)_{\text{obs}}.\end{aligned}$$

When $v_C = w_C = v_A = w_T = v_G = 1$, $r_C^A = r_C^G = r_T^A = 0$ and $r_T^G = r$ (as in JC+CpG), $(\kappa_{RN}^N)_{\text{obs}} = (\kappa_C^N)_{\text{obs}}$.

The expression for the observed quantity $(v_C^N)_{\text{obs}}$ is unchanged between JC+CpG and RN+YpR because lemma 3.3.1 holds in the general case.

Once again, Slutsky's lemma for the observed quantities $(\kappa_{RN}^N)_{\text{obs}}$ and $(v_{RN}^N)_{\text{obs}}$ is the key to state theorem 3.8.1 below, which is a consequence of proposition 3.3.5.

Theorem 3.8.1. *Assume that the ancestral sequence is at stationarity and that conjecture 3.3.4 holds. Then, when $N \to +\infty$,*

$$\kappa_{\text{obs}}^{RN} \sqrt{N / v_{\text{obs}}^{RN}} (T_C - t)$$

converges in distribution to the standard normal law. An asymptotic confidence interval at level ε for t is

$$\left[(T_C^N) - \frac{z(\varepsilon)}{(\kappa_{RN}^N)_{\text{obs}}} \sqrt{\frac{(v_{RN}^N)_{\text{obs}}}{N}}, (T_C^N) + \frac{z(\varepsilon)}{(\kappa_{RN}^N)_{\text{obs}}} \sqrt{\frac{(v_{RN}^N)_{\text{obs}}}{N}} \right],$$

where $z(\varepsilon)$ denotes the unique real number such that $\mathbb{P}(|Z| \geqslant z(\varepsilon)) = \varepsilon$ with Z a variable with standard normal law.

As in JC+CpG, the estimator (T_C^N) is defined implicitly for RN+YpR. However, in JC+CpG, we provided an explicit expression for $(C,C)(t)$, but, even for Kimura + CpG, that is with parameters

$$v_A = v_T = v_C = v_G = 1, \quad w_A = w_T = w_C = w_G = \kappa,$$
$$r_C^A = r_G^T = r_G^C = r_T^A = r_C^G = r_A^T = 0, \quad \text{and} \quad r_T^G = r_A^C = r,$$

where κ denotes the transition/transversion rate, we find difficult to provide an explicit expression for $(C,C)(t)$. Fortunately, numerical methods allow to compute a closed form of the theoretical solution of the differential linear system, and consequently it is possible to solve equation $(C,C)(t) = (C,C)_{\text{obs}}$ with numerical methods.

3.9 Evolution of $(C,C)(t)$ in RN+YpR

We base our description of the method in the general RN+YpR model on the encoding of dinucleotides as

$$\mathscr{B} = \{R,T,C\} \times \{Y,G,A\},$$

and we show that its evolution is autonomous.

We now define a 9×9 matrix m, indexed by $\mathscr{B} \times \mathscr{B}$, hence uv and xy are generic elements of $\mathscr{B}$.

Let v_R and v_Y denote

$$v_R = v_A + v_G, \quad v_Y = v_T + v_C.$$

Then,

$$\begin{aligned}
m(uv,xy) &= 0, \quad \text{if} \quad u \neq x \quad \text{and} \quad v \neq y;\\
m(Rx,ux) &= v_u, \quad \text{if} \quad x \in \{Y,G,A\} \quad \text{and} \quad u \in \{C,T\};\\
m(ux,Rx) &= v_R, \quad \text{if} \quad x \in \{Y,G,A\} \quad \text{and} \quad u \in \{C,T\};\\
m(Ru,Rv) &= w_v, \quad \text{if} \quad \{u,v\} = \{A,G\};\\
m(xY,xu) &= v_u, \quad \text{if} \quad x \in \{R,C,T\} \quad \text{and} \quad u \in \{A,G\};\\
m(xu,xY) &= v_R, \quad \text{if} \quad x \in \{R,C,T\} \quad \text{and} \quad u \in \{A,G\};\\
m(uY,vY) &= w_v, \quad \text{if} \quad \{u,v\} = \{T,C\};\\
m(xu,xv) &= w_v + r_v^x, \quad \text{if} \quad \{u,v\} = \{A,G\} \quad \text{and} \quad x \in \{T,C\};\\
m(ux,vx) &= w_v + r_v^x, \quad \text{if} \quad \{u,v\} = \{C,T\} \quad \text{and} \quad x \in \{A,G\}.
\end{aligned}$$

Then, for every uv and xy in $\mathscr{B}$, the function (uv,xy) satisfies the linear differential equation

$$(uv,xy)'(t) = \sum_{(w,z)\in\mathscr{B}} m(wz,xy)(uv,wz)(t).$$

Hence quantities such as $(C,C)(t)$ can be computed provided one computes the exponential of the rate-matrix, and that quantities such as $(C,C)'(t)$ have computable explicit expressions in terms of frequencies expressed in the reduced dinucleotide-alphabet $\mathscr{B}$.

3.10 Simulations

As a support to the conjecture that $t \mapsto (C,C)(t)$ always defines a diffeomorphism for models in the class RN+YpR, we performed some simulations whose aim is to draw an approximation of the graph of $t \mapsto (C,C)(t)$. Every set of parameters we tested supports the conjecture. We present the way we proceeded.

We use the encoding of dinucleotides by the reduced alphabet $\mathscr{B}$. For every xy in $\mathscr{B}$, the function $t \mapsto (C*,xy)(t)$ satisfies the linear differential equation

$$(C*,xy)'(t) = \sum_{(u,v)\in\mathscr{B}} m(uv,xy)(C*,uv)(t), \tag{3.10.1}$$

where $m(uv,xy)$ corresponds to the 9×9 matrix introduced in section 3.9. This matrix depends on 16 parameters: 8 parameters for the substitution rates which characterize the independent evolution, and 8 parameters for the specific rates due to the influence mechanism.

To solve the linear differential system defined by the generic equation (3.10.1), we need to know the initial value of every function $(C*,xy)$ when the initial sequence is at stationarity. Of course, $(C*,xy)(0) = 0$ if $x \neq C$ and $(C*,Cy)(0) = (Cy)_*$. To compute the stationary frequencies, we solve the linear system

$$(xy)_* = \sum_{(u,v)\in\mathscr{B}} m(uv,xy)(uv)_*. \tag{3.10.2}$$

Since $(C,C)(t)$ is a linear combination of terms of the form $(C*,xy)(t)$, that is,

$$(C,C)(t) = (C*,CA)(t) + (C*,CY)(t) + (C*,CG)(t),$$

we obtain an expression of $(C,C)(t)$ if we solve the linear differential system given by the generic equation (3.10.1). We use numerical methods with `Maple` to solve systems (3.10.2) and (3.10.1). Finally, we draw the approximation of the graph of $t \mapsto (C,C)(t)$.

We present below a table containing the range of parameter values that we explored, and the corresponding evolution models. Then, we present the figures performed for each set of parameters, which represent approximations of the theoretical function $t \mapsto (C,C)(t)$.

K80 is a shorthand for Kimura 80 [Kim80], HKY85 for Hasegawa *et al.* 1984, 1985 [HKY85], and TN93 for Tamura and Nei 1993 [TN93].

3.10.1 Range of parameter values explored

Model Simulation	v_A r_A^C	v_T r_C^A	v_C r_G^C	v_G r_C^G	w_A r_T^G	w_T r_G^T	w_C r_T^A	w_G r_A^T
JC+CpG	1	1	1	1	1	1	1	1
Simulation 1	10	0	0	0	10	0	0	0
K80+CpG	1	1	1	1	3	3	3	3
Simulation 2	10	0	0	0	10	0	0	0
K80+CpG	1	1	1	1	0.3	0.3	0.3	0.3
Simulation 3	10	0	0	0	10	0	0	0
K80+YpR	1	1	1	1	0.3	0.3	0.3	0.3
Simulation 4	10	10	10	10	10	10	10	10
K80+YpR	1	1	1	1	3	3	3	3
Simulation 5	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
HKY85+CpG	0.7	0.7	0.3	0.3	2.1	2.1	0.9	0.9
Simulation 6	10	0	0	0	10	0	0	0
HKY85+CpG	0.3	0.3	0.7	0.7	0.9	0.9	2.1	2.1
Simulation 7	10	0	0	0	10	0	0	0
HKY85+YpR	0.7	0.7	0.3	0.3	2.1	2.1	0.9	0.9
Simulation 8	10	10	10	10	10	10	10	10
HKY85+YpR	0.7	0.7	0.3	0.3	2.1	2.1	0.9	0.9
Simulation 9	10	1	3	7	10	1	3	7
HKY85+YpR	0.7	0.7	0.3	0.3	2.1	2.1	0.9	0.9
Simulation 10	1	0.1	0.3	0.7	1	0.1	0.3	0.7
TN93+CpG	0.3	0.4	0.1	0.2	1.5	1.2	0.3	1.0
Simulation 11	10	0	0	0	10	0	0	0
TN93+YpR	0.3	0.4	0.1	0.2	1.5	1.2	0.3	1.0
Simulation 12	10	1	3	7	10	1	3	7
TN93+YpR	0.3	0.4	0.1	0.2	1.5	1.2	0.3	1.0
Simulation 13	1	0.1	0.3	0.7	1	0.1	0.3	0.7
RN+CpG	1	2	0.5	4	2	3	4	8
Simulation 14	10	0	0	0	10	0	0	0
RN+YpR	1	2	0.5	4	2	3	4	8
Simulation 15	10	1	3	7	10	1	3	7
RN+YpR	1	2	0.5	4	2	3	4	8
Simulation 16	1	0.1	0.3	0.7	1	0.1	0.3	0.7
RN+YpR	1	2	0.5	4	2	3	4	8
Simulation 17	1	0.1	0.3	0.7	1.5	0.2	0.6	1.4

3.10.2 Figures performed on `Maple`

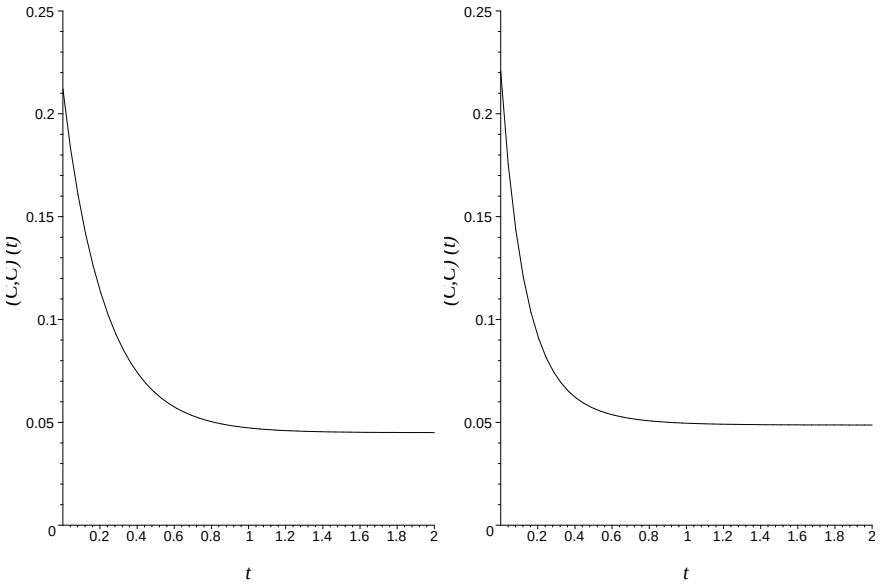


Figure 3.6: Simulations 1 and 2

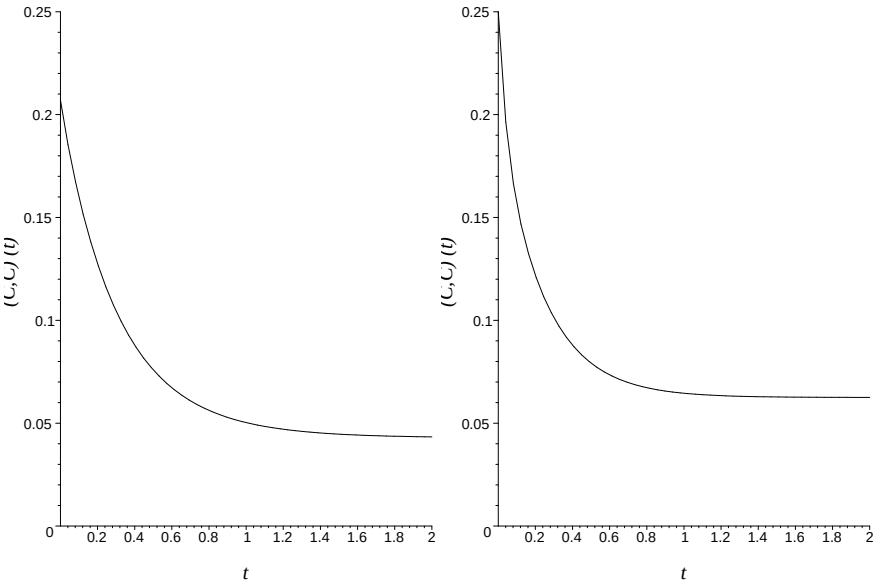


Figure 3.7: Simulations 3 and 4

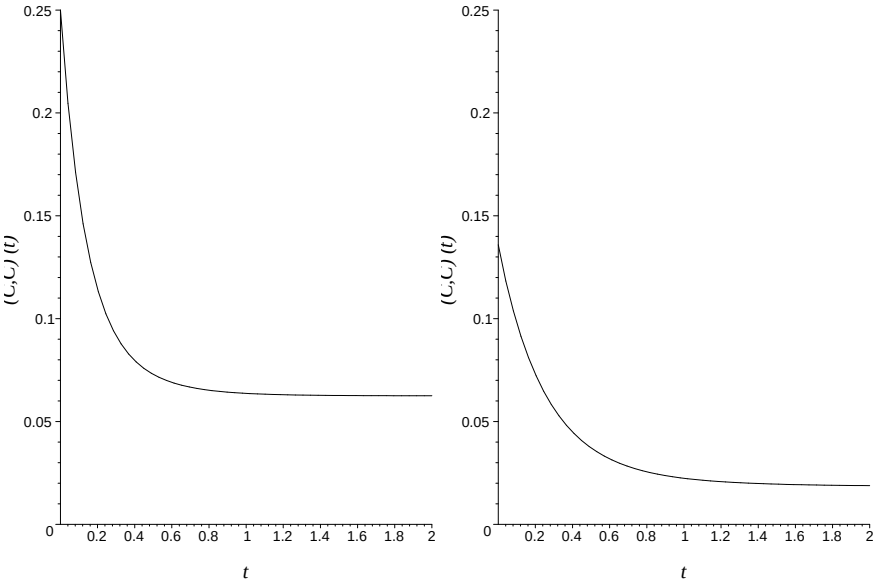


Figure 3.8: Simulations 5 and 6

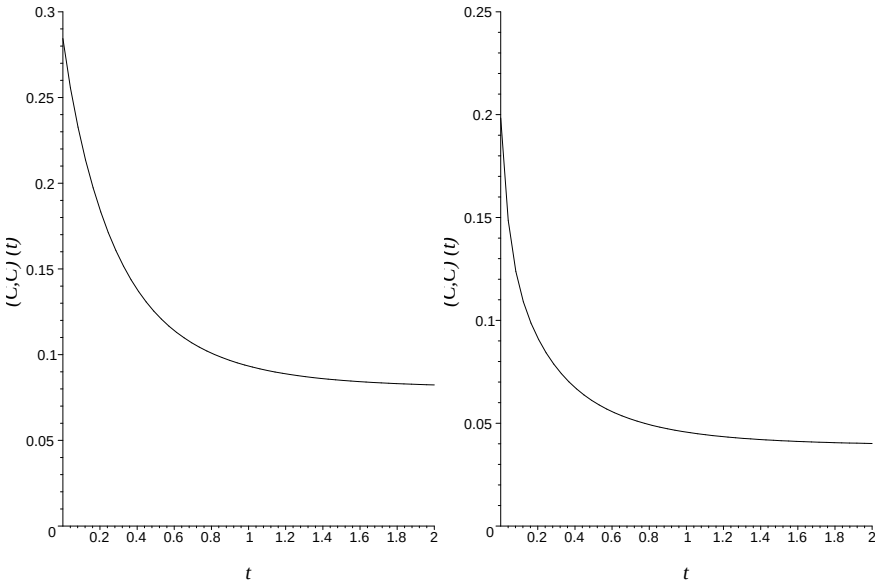


Figure 3.9: Simulations 7 and 8

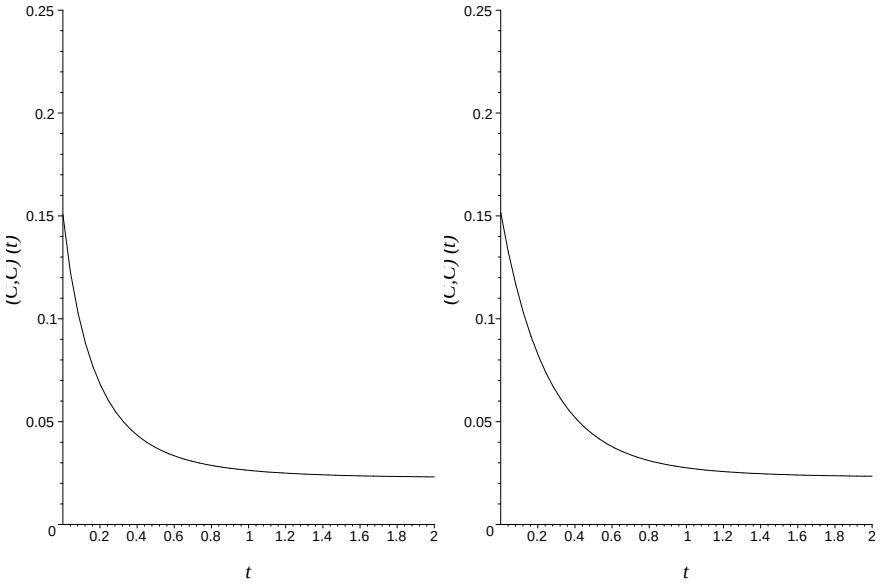


Figure 3.10: Simulations 9 and 10

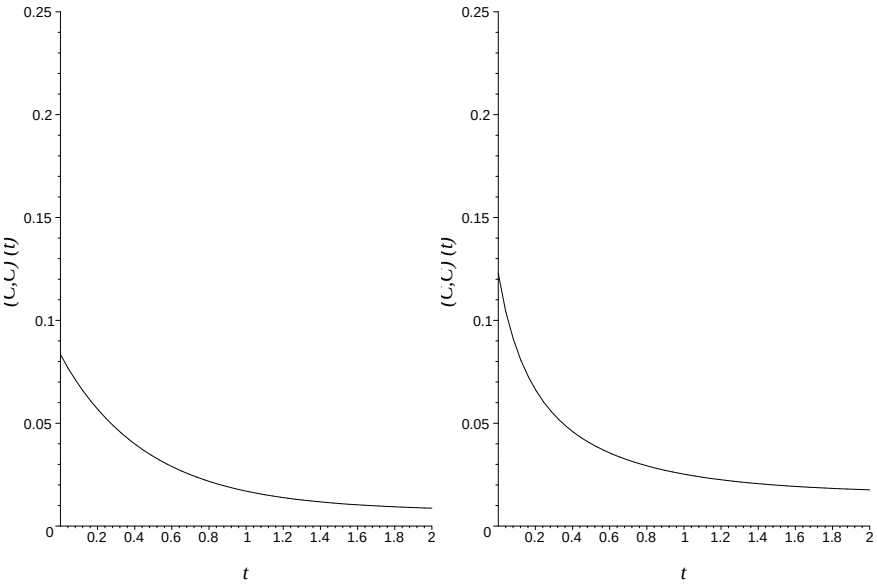


Figure 3.11: Simulations 11 and 12

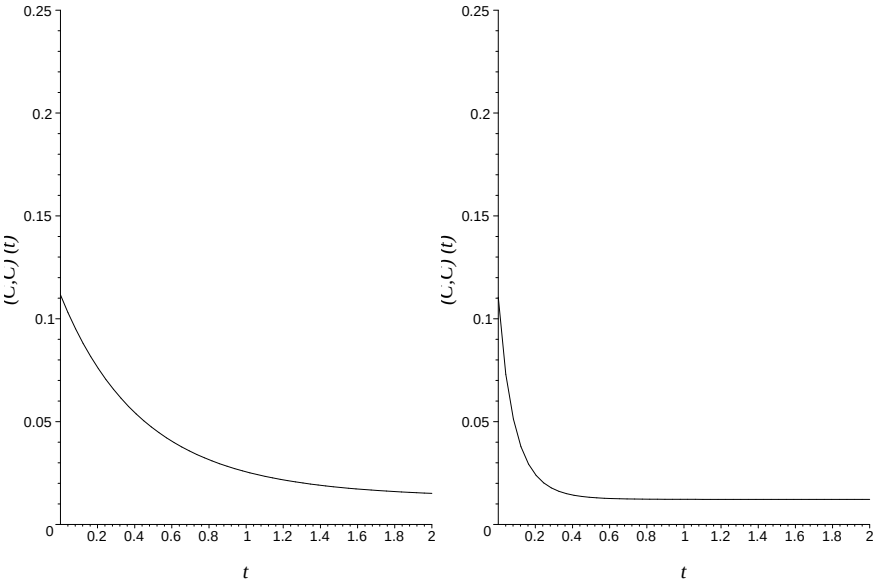


Figure 3.12: Simulations 13 and 14

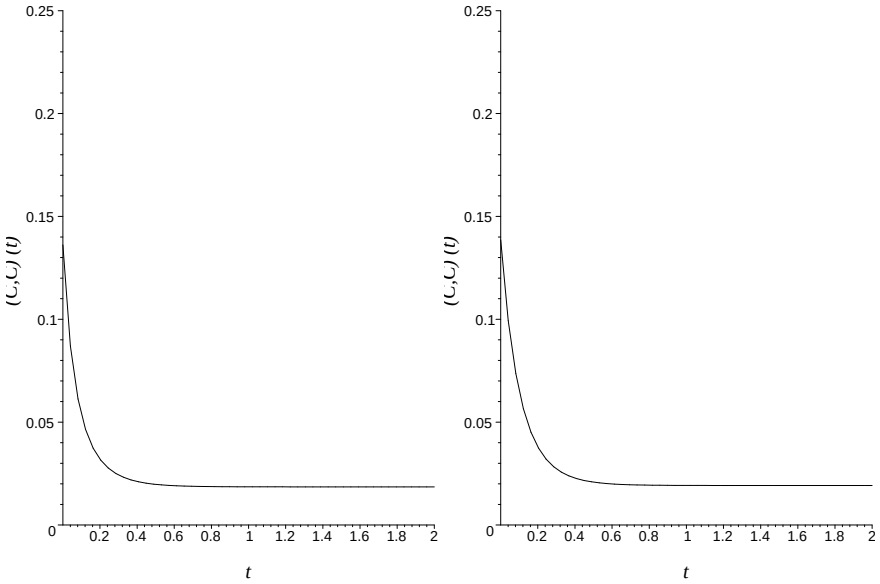


Figure 3.13: Simulations 15 and 16

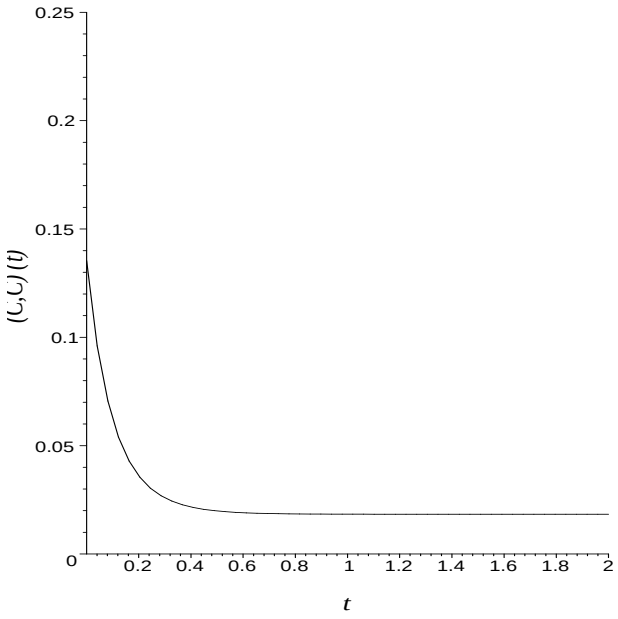


Figure 3.14: Simulation 17

Chapter 4

Priors for the Bayesian star paradox

In phylogenetics, a particular resolved tree can be highly supported even when the data is generated by an unresolved star tree. This unfortunate aspect of the Bayesian approach to phylogeny reconstruction is called the *star paradox*. Recent studies highlight that the paradox can occur in the simplest setting, namely, for an unresolved rooted tree on three taxa and two states, see Yang and Rannala [YR05] and Lewis et al. [HLH05] for example. Kolaczkowski and Thornton presented in [KT06] some simulations and suggested that artifactual high posteriors for a particular resolved tree might disappear for very long sequences. Previous simulations in Yang and Rannala's paper were plagued by numerical problems, which left unknown the nature of the limiting distribution on posterior probabilities. For an introduction to the Bayesian approach to phylogeny reconstruction see chapter 5 of Yang [Yan06].

The statistical question which supports the star paradox is whether the Bayesian posterior distribution of the resolutions of a star tree becomes uniform when the length of the sequence tends to infinity, that is, in the case of three taxa, whether the posterior distribution of each resolution converges to $1/3$. In a recent paper, Steel and Matsen [SM07] disprove this, thus ruining Kolaczkowski and Thornton's hope, for a specific class of branch length priors which they call *tame* (a prior is tame if its distribution has a smooth joint probability density function that is bounded and everywhere non zero). More precisely, Steel and Matsen show that, for every tame prior and every fixed $\varepsilon > 0$, the posterior probability of any of the three possible trees stays above $1 - \varepsilon$ with non vanishing probability when the length of the sequence goes to infinity. This result had been taken account by Yang in [Yan07] and reinforced by theoretical results on the posterior probabilities by Susko in [Sus08].

Our main result is that Steel and Matsen's conclusion holds for a wider class of

priors, possibly not continuous, which we call *tempered* and define in section 4.2. Recall that Steel and Matsen consider smooth priors, whose densities satisfy some regularity conditions.

The chapter is organized as follows. In section 4.1, we describe the Bayesian framework of the star paradox. In section 4.2, we define the class of tempered priors for branch lengths and we state our main result. In section 4.3, we state an extension of a technical lemma due to Steel and Matsen, which allows us to extend their result. In section 4.4, we prove our main result. Section 4.5 is devoted to the proofs of intermediate results. In section 4.6, we prove that every tame prior, in Steel and Matsen's sense, is tempered, in the sense of this chapter, and we provide examples of tempered, but not tame, prior distributions. Finally, in section 4.7, we prove the extension of Steel and Matsen's technical lemma.

4.1 Bayesian framework for rooted trees on three taxa

Consider three taxa, encoded by the set $\tau = \{1,2,3\}$, with two possible states. Phylogenies on τ are supported by one of the four following trees: the star tree R_0 on three taxa and, for every taxon $i \in \tau$, the tree R_i such that i is the outlier, hence

$$R_1 = (1,(2,3)), \quad R_2 = (2,(1,3)), \quad R_3 = (3,(1,2)).$$

The phylogeny based on R_0 is specified by the common length of its three branches, denoted by t. For each $i \in \tau$, the phylogeny based on R_i is specified by a couple of branch lengths $(t_\mathrm{e}, t_\mathrm{i})$, where t_e denotes the external branch length and t_i the internal branch length, see figure 4.1.

For instance, in the phylogeny based on R_1, the divergence of taxa 2 and 3 occurred t_e units of time ago and the divergence of taxon 1 from taxa 2 and 3 occurred $t_\mathrm{i} + t_\mathrm{e}$ units of time ago.

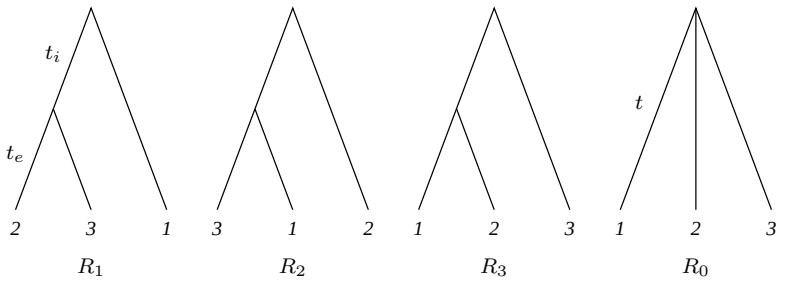


Figure 4.1: The four rooted trees for three species.

Four site patterns can occur on τ: s_0 denotes the pattern such that a given site coincides in the three taxa and, for every $i \in \tau$, s_i denotes the pattern such that a

given site coincide in the two other taxa and is different in taxon i. In other words, if one writes the site patterns in taxa 1, 2 and 3 in this order and x and y for any two different characters,

$$s_0 = xxx, \quad s_1 = yxx, \quad s_2 = xyx, \quad \text{and} \quad s_3 = xxy.$$

Let π denote the set of site patterns. As explained above, in the specific case of three taxa and two states evolving in a Jukes-Cantor model, one can choose $\pi = \tau \cup \{0\}$, in effect using $i \in \pi$ as a shorthand for the collection of site patterns encoded by s_i. Assume that the counting of site pattern i is n_i. Then $n = n_0 + n_1 + n_2 + n_3$ is the total length of the sequences and, in the independent Jukes-Cantor model considered in this chapter, the quadruple (n_0, n_1, n_2, n_3) is a sufficient statistics of the sequence data. We use $n_{0:3}$ to denote any quadruple (n_0, n_1, n_2, n_3) of nonnegative integers such that $|n_{0:3}| = n_0 + n_1 + n_2 + n_3 = n \geqslant 1$.

We assume that the sequences evolve according to a continuous-time Markov process with equal substitution rates 1 between the two characters.

For every $i \in \pi$ and every couple of branch lengths $(t_\mathrm{e}, t_\mathrm{i})$, let $p_i(t_\mathrm{e}, t_\mathrm{i})$ denote the probability that site pattern s_i occurs on tree R_1 with branch lengths $(t_\mathrm{e}, t_\mathrm{i})$. Standard computations provided by Yang and Rannala show that

$$\begin{aligned}
4p_0(t_\mathrm{e}, t_\mathrm{i}) &= 1 + \mathrm{e}^{-4t_\mathrm{e}} + 2\mathrm{e}^{-4(t_\mathrm{i}+t_\mathrm{e})}, \\
4p_1(t_\mathrm{e}, t_\mathrm{i}) &= 1 + \mathrm{e}^{-4t_\mathrm{e}} - 2\mathrm{e}^{-4(t_\mathrm{i}+t_\mathrm{e})}, \\
4p_2(t_\mathrm{e}, t_\mathrm{i}) &= 4p_3(t_\mathrm{e}, t_\mathrm{i}) = 1 - \mathrm{e}^{-4t_\mathrm{e}}.
\end{aligned}$$

Let $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ denote a couple of positive random variables representing the branch lengths $(t_\mathrm{e}, t_\mathrm{i})$. Let $N_{0:3} = (N_0, N_1, N_2, N_3)$ denote a random variable representing the counts of sites patterns $n_{0:3} = (n_0, n_1, n_2, n_3)$. We often write N for $N_{0:3}$.

4.2 The star tree paradox

Assuming that every taxon in τ evolved from a common ancestor, the aim of phylogeny reconstruction is to compute the most likely tree R_i. To do so, in the Bayesian approach, one places prior distributions on the trees R_i and on their branch lengths $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$.

4.2.1 Main result

Let $\mathbb{P}(N = n_{0:3} | R_i, \mathfrak{T})$ denote the probability that $N = n_{0:3}$ assuming that the data is generated along the tree R_i conditionally on the branch lengths $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$. One may consider R_1 only since, for every $n_{0:3}$, the symmetries of the setting yield the relations

$$\mathbb{P}(N = n_{0:3} | R_2, \mathfrak{T}) = \mathbb{P}(N = (n_0, n_2, n_3, n_1) | R_1, \mathfrak{T}),$$

and

$$\mathbb{P}(N = n_{0:3}|R_3, \mathfrak{T}) = \mathbb{P}(N = (n_0, n_3, n_1, n_2)|R_1, \mathfrak{T}).$$

Notation 4.2.1. *For every $i \in \tau$, let $\tau_i = \tau \setminus \{i\}$. For every $i \in \pi$, let P_i denote the random variable*

$$P_i = p_i(\mathfrak{T}) = p_i(T_\mathrm{e}, T_\mathrm{i}).$$

For every $i \in \tau$ and every $n_{0:3}$, let $\Pi_i(n_{0:3})$ denote the random variable

$$\Pi_i(n_{0:3}) = P_0^{n_0} P_1^{n_i} P_2^{n_j+n_k}, \quad \textit{with} \quad \{i, j, k\} = \tau.$$

We recall that $P_2 = P_3$ and we note that, if $|n_{0:3}| = n_0 + n_1 + n_2 + n_3 = n$ with $n \geqslant 1$, for every $i \in \tau$,

$$\Pi_i(n_{0:3}) = P_0^{n_0} P_1^{n_i} P_2^{n-n_0-n_i}.$$

Fix $n_{0:3}$ and assume that $|n_{0:3}| = n_0 + n_1 + n_2 + n_3 = n$ with $n \geqslant 1$. For every $i \in \tau$, the posterior probability of R_i conditionally on $N = n_{0:3}$ is

$$\mathbb{P}(R_i|N = n_{0:3}) = \frac{n!}{n_0!n_1!n_2!n_3!} \frac{1}{\mathbb{P}(N = n_{0:3})} \mathbb{E}(\Pi_i(n_{0:3})).$$

Thus, for every i and $j \in \tau$,

$$\frac{\mathbb{P}(R_i|N = n_{0:3})}{\mathbb{P}(R_j|N = n_{0:3})} = \frac{\mathbb{E}(\Pi_i(n_{0:3}))}{\mathbb{E}(\Pi_j(n_{0:3}))}.$$

Definition 4.2.2. *For every $\varepsilon > 0$ and every $i \in \tau$, let $\mathfrak{N}_i^\varepsilon$ denote the set of $n_{0:3}$ such that, for both indices $j \in \tau$ such that $j \neq i$,*

$$\mathbb{E}(\Pi_i(n_{0:3})) \geqslant (2/\varepsilon)\, \mathbb{E}(\Pi_j(n_{0:3})).$$

For every $i \in \tau$ and $n_{0:3} \in \mathfrak{N}_i^\varepsilon$,

$$\mathbb{P}(R_i|N = n_{0:3}) \geqslant 1 - \varepsilon,$$

which means that the posterior probability of tree R_i among the three possible trees is highly supported.

Recall that, under hypothesis R_0 and for a tame prior distribution on $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$, that is, with a smooth joint probability density function that is bounded and everywhere non zero, Steel and Matsen prove that, for every $i \in \tau$, $\mathbb{P}(N \in \mathfrak{N}_i^\varepsilon)$ does not go to 0 when the sequence length n goes to infinity, and consequently that the posterior probability $\mathbb{P}(R_i|N)$ can be close to 1 even when the sequence length n is large.

We prove the same result for tempered prior distributions of $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$, which we now define.

Notation 4.2.3. *(1) For every $s \in [0,1]$ and $z \in [0,3]$, let*

$$G_s(z) = \mathbb{P}\left(e^{-4T_e}(1-e^{-4T_i}) \leqslant s \,|\, e^{-4T_e}(1+2e^{-4T_i}) = z\right).$$

(2) For every positive t and every $i \in \pi$, let q_i denote the probability that site pattern s_i occurs on tree R_0, hence

$$4q_0 = 4p_0(0,t) = 1+3e^{-4t}, \quad 4q_1 = 4q_2 = 4q_3 = 1-e^{-4t}.$$

(3) Let ℓ_t denote a positive real number such that $1 < 4q_0 - \ell_t$ and $4q_0 + \ell_t < 4$, for instance $\ell_t = 3e^{-4t}(1-e^{-4t})$. Let I and I_t denote the intervals

$$I = [0,3], \quad I_t = [4q_0 - 1 - \ell_t, 4q_0 - 1 + \ell_t] \subset]0,3[.$$

(4) For every positive t and integer n, let

$$Q_n(t) = \mathbb{P}\left(T_i \leqslant 1/n, t \leqslant T_e \leqslant t + 1/n\right).$$

Definition 4.2.4 (Tempered priors)**.** *The distribution of $\mathfrak{T} = (T_e, T_i)$ is tempered if the following two conditions hold.*

1. *For every t, there exists $s_0 \in]0,1]$, an interval I_t around $4q_0 - 1$, bounded functions $(F_i)_{i=0}^{k-1}$, positive numbers α and κ, and real numbers $(\varepsilon_i)_{i=0}^{k}$ such that*

   $$0 = \varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_{k-1} \leqslant 2 < \varepsilon_k,$$

   and such that for every $s \in [0, s_0]$ and every $z \in I_t$,

   $$\left| G_s(z) - \sum_{i=0}^{k-1} F_i(z) s^{\alpha+\varepsilon_i} \right| \leqslant \kappa s^{\alpha+\varepsilon_k}.$$

2. *For every positive t, $n^{-1}\log Q_n(t) \to 0$ when $n \to \infty$.*

We detail the properties involved in definition 4.2.4 and provide examples of tempered priors in subsection 4.2.2 below.

We now state our main result, which is the extension of Steel and Matsen's result to our more general setting.

Theorem 4.2.5. *Consider sequences of length n generated by a star tree R_0 on 3 taxa with strictly positive edge length t. Let N be the resulting data, summarized by site pattern counts. Consider any prior on the three resolved trees (R_1, R_2, R_3) and a tempered prior distribution on their branch lengths $\mathfrak{T} = (T_e, T_i)$.*
Then, for every $i \in \tau$, for every positive ε, there exists a positive δ such that, when n is large enough,

$$\mathbb{P}\left((\mathbb{P}(R_i|N) \geqslant 1-\varepsilon) \geqslant \delta.\right.$$

We prove theorem 4.2.5 in section 4.4.

4.2.2 Motivation and intuitive understanding of definition 4.2.4

In definition 4.2.4, condition (2) is easy to describe, to illustrate and to check, but condition (1) might be more difficult to grasp. Condition (1) involves a Taylor expansion around $s = 0$ of the function $s \mapsto G_s(z)$, where the Taylor coefficients depend on z. The Taylor expansion comes from a technical result given in proposition 4.3.2, which concerns the asymptotics of $1 - M_{t+1}/M_t$, where $M_t = \mathbb{E}(V^t)$ with V a $[0,1]$ random variable, and depends on the behaviour of the distribution function of V around 1. Theorem 4.2.5 relies on this technical result and this explains the Taylor expansion.

Now, the main difficulty lies in the following problem: given a prior, how to check if the prior is tempered or not? To wit, to check that a Taylor expansion holds for the function $s \mapsto G_s(z)$, one needs to state its expression, and compute a conditional expectation involving two random variables, which can be fastidious. That is why we now present some explicit examples of tempered priors. We begin with the following result.

Proposition 4.2.6. *Assume that $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ has a smooth joint probability density ω which is bounded and everywhere non zero. Then the distribution of $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ is tempered.*

As a consequence, every tame prior fulfills the hypothesis of proposition 4.2.6, hence every tame prior is tempered, as claimed in the introduction. This case includes the exponential priors discussed in [YR05]. We prove proposition 4.2.6 in section 4.6.

However some tempered priors are not tame, as illustrated in the following example where Steel and Matsen's condition fails.

Definition 4.2.7. *Fix $a > 0$ and $b > 0$. Let (t_n), (y_n) and (r_n) denote sequences of positive numbers, indexed by $n \geqslant 1$, and r a positive number, defined by the formulas*

$$t_n = n^{-a}, \quad y_n = 1 + 2\mathrm{e}^{-4t_n} \quad r_n = y_n\,[n^{-b} - (n+1)^{-b}], \quad r = \sum_{n \geqslant 1} r_n.$$

Proposition 4.2.8. *Choose positive parameters a and b such that $3a < b$ and $3a < 1$. Assume that T_i is a discrete random variable such that, for every $n \geqslant 1$,*

$$\mathbb{P}(T_\mathrm{i} = t_n) = r_n/r.$$

Assume that T_e is a continuous random variable, independent of T_i, with exponential law of parameter 4, that is, with density $4\,\mathrm{e}^{-4t}$ on $t \geqslant 0$ with respect to the Lebesgue measure.

Then, the distribution of $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ is not tame but it is tempered, for the parameters

$$k = 3, \quad \alpha = b/a, \quad \varepsilon_1 = 1, \quad \varepsilon_2 = 2, \quad \varepsilon_3 = 3,$$

and some explicit functions F_0, F_1 and F_2.

Since the distribution of T_i is an accumulation of Dirac distributions, the prior distribution of $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ cannot be tame.

Yet, the fact that the prior distribution is tempered does not come from the fact that the distribution of T_i is discrete. For a degenerate example, if $T_\mathrm{i} = 0$ almost surely, then $G_s(z) = 1$ for every $s > 0$, and this function does not have a Taylor expansion around zero whose first term is a positive power of s. Note that in this particular case, the Bayesian star paradox does not occur.

However, under the conditions of proposition 4.2.8, one can compute a Taylor expansion of $G_s(z)$ around zero which fulfills condition (1) of definition 4.2.4. We prove this in section 4.6.

We provide below some examples of less ill-behaved distributions which are tempered but not tame, and one example of distribution which does not fulfill condition (1), hence is not tempered.

Proposition 4.2.9. *Assume that T_e is a continuous random variable, with exponential law of parameter 4, that is, with density $4\,\mathrm{e}^{-4t}$ on $t \geqslant 0$ with respect to the Lebesgue measure. Assume that T_i is a random variable independent of T_e.*

(i) If the distribution of T_i is uniform on $[0, \tau]$, with $\tau > 0$, the distribution of $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ is tempered but not tame.

(ii) If the distribution of T_i has density $\kappa t_\mathrm{i}^{\kappa-1}$ on the interval $[0,1]$, for a given κ in $(0,1)$, the distribution of $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ is tempered but not tame.

(iii) If the distribution of T_i has density $-\log(t_\mathrm{i})$ on the interval $[0,1]$, the distribution of $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ does not fulfill condition (1) of definition 4.2.4.

We prove proposition 4.2.9 in section 4.6.

4.3 Extension of Steel and Matsen's lemma

The Bayesian star paradox due to Steel and Matsen relies on a technical result which we slightly rephrase as follows. For every nonnegative real t and every $[0,1]$ valued random variable V, introduce

$$M_t = \mathbb{E}(V^t), \quad R_t = 1 - \frac{M_{t+1}}{M_t} = \frac{\mathbb{E}\left(V^t(1-V)\right)}{\mathbb{E}(V^t)}.$$

Proposition 4.3.1 (Steel and Matsen's lemma)**.** *Let $0 \leqslant \eta < 1$ and $B > 0$. There exists a finite K, which depends on η and B only, such that the following holds. For every $[0,1]$ valued random variable V with a smooth probability density function f such that $f(1) > 0$ and $|f'(v)| \leqslant B f(1)$ for every $\eta \leqslant v \leqslant 1$, and for every integer $k \geqslant K$,*

$$2kR_k \geqslant 1.$$

Indeed the asymtotics of R_k when k is large depends on the distribution of V around 1. We prove in the following proposition that the conclusion of Steel and Matsen's lemma above holds for a wider class of random variables.

Proposition 4.3.2. *Let V a random variable on $[0,1]$. Suppose that there exists real numbers $0 \leqslant v_0 < 1$, $\alpha > 0$, $(\varepsilon_i)_{i=0}^n$ and $(\gamma_i)_{i=0}^n$, such that*

$$0 = \varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_{n-1} \leqslant 1 < \varepsilon_n,$$

and, for every $v_0 \leqslant v \leqslant 1$,

$$\left| \mathbb{P}(V \geqslant v) - \sum_{i=0}^{n-1} \gamma_i (1-v)^{\alpha+\varepsilon_i} \right| \leqslant \gamma_n (1-v)^{\alpha+\varepsilon_n}.$$

Then there exists a finite $\tau(\gamma_{0:n})$, which depends continuously on $\gamma_{0:n}$, such that for every $t \geqslant \tau(\gamma_{0:n})$,

$$2tR_t \geqslant \alpha.$$

Remark 4.3.3. *We insist on the fact that $\tau(\gamma_{0:n})$ depends continuously on $\gamma_{0:n}$. To wit, in the proof of proposition 4.5.6, we apply proposition 4.3.2 with bounded functions of z. This means that for every $z \in I_t$, one gets a number τ which depends on z through the bounded functions such the control on the distribution of V holds. The continuity of τ ensures that there exists a number independent of z such that proposition 4.5.6 holds.*

Remark 4.3.4. *If one computes a Taylor expansion of the function $v \mapsto \mathbb{P}(V \geqslant v)$ at $v = 1^-$ under the conditions of Steel and Matsen's lemma, one can see that conditions of proposition 4.3.2 hold. It follows that proposition 4.3.2 is an extension of Steel and Matsen's lemma.*

We prove proposition 4.3.2 in section 4.7. The proof of theorem 4.2.5 relies on this technical proposition.

4.4 Sketch of proof of theorem 4.2.5

This section is devoted to the sketch of the proof of theorem 4.2.5. We use the definitions below. Note that the set $F_c^{(n)}$ is not the set introduced by Steel and Matsen. For a technical reason in the proof of proposition 4.4.2 stated below, we had to modify their definition. Note however that propositions 4.4.2 and 4.4.3 below are adaptations of ideas in Steel and Matsen's paper.

Notation 4.4.1. *For every i in π, let Δ_i denote the function defined as follows. For every nonnegative integers $n_{0:3} = (n_0, n_1, n_2, n_3)$ such that $|n_{0:3}| = n_0 + n_1 + n_2 + n_3 = n$ with $n \geqslant 1$,*

$$\Delta_0(n_{0:3}) = \frac{n_0 - q_0 n}{\sqrt{n}},$$

and, for every $i \in \tau$,

$$\Delta_i(n_{0:3}) = \frac{n_i - 1/3(n-n_0)}{\sqrt{n}}.$$

For every $c > 1$, introduce

$$F_c^{(n)} = \{n_{0:3}\,;\, |n_{0:3}| = n,\, -2c \leqslant \Delta_2(n_{0:3}) \leqslant -c,\, -2c \leqslant \Delta_3(n_{0:3}) \leqslant -c,\, -c \leqslant \Delta_0(n_{0:3}) \leqslant 0\}.$$

For every $i \in \tau$ and every positive η, let A_η^i denote the event

$$A_\eta^i = \left\{\forall j \in \tau_i,\, \mathbb{E}(\Pi_i(N)\,|N) \geqslant \eta \mathbb{E}(\Pi_j(N)\,|N)\right\}.$$

Since $\Delta_1 + \Delta_2 + \Delta_3 = 0$, every $n_{0:3}$ in $F_c^{(n)}$ is such that $2c \leqslant \Delta_1(n_{0:3}) \leqslant 4c$. We note that $F_c^{(n)}$ is not symmetric about τ and gives a preference to 1. That is why we only deal with A_η^1 in the following proof. To deal with A_η^i, it suffices to modify the definition of $F_c^{(n)}$.

From the reasoning in section 4.2, it suffices to prove that for every positive η, there exists a positive δ such that, when n is large enough,

$$\mathbb{P}\left(A_\eta^1\right) \geqslant \delta.$$

Suppose that one generates $n \geqslant 1$ sites on the star tree R_0 with given branch length t and let N be the counts of site patterns defined in section 4.1, hence $N_0 + N_1 + N_2 + N_3 = n$.

When n is large enough, central limit estimates show that the probability of the event $\left\{N \in F_c^{(n)}\right\}$ is uniformly bounded from below, say by $\delta > 0$. Hence,

$$\mathbb{P}\left(A_\eta^1\right) \geqslant \delta \mathbb{P}\left(A_\eta^1 \middle| N \in F_c^{(n)}\right)$$

We wish to prove that there exists a positive α independent of c such that for n large enough and for every $n_{0:3} \in F_c^{(n)}$, and for every $j \in \tau_1$,

$$\mathbb{E}(\Pi_1(n_{0:3})) \geqslant c^2 \alpha\, \mathbb{E}(\Pi_j(n_{0:3})).$$

This follows from the two propositions below, adapted from Steel and Matsen's paper.

Proposition 4.4.2. *Fix t and assume that $n_{0:3} \in F_c^{(n)}$. Then, when n is large enough, for every $j \in \tau_1$,*

$$\mathbb{E}(\Pi_j(n_{0:3})\,|\,4P_0 - 1 \in I_t) \geqslant \mathbb{E}(\Pi_j(n_{0:3})\,|\,4P_0 - 1 \notin I_t).$$

Proposition 4.4.3. *Fix t and assume that $n_{0:3} \in F_c^{(n)}$. Then, there exists a positive α, independent of c, such that for every $z \in I_t$, and for every $j \in \tau_1$,*

$$\mathbb{E}(\Pi_1(n_{0:3})\,|\,4P_0 - 1 = z) \geqslant c^2 \alpha \mathbb{E}(\Pi_j(n_{0:3})\,|\,4P_0 - 1 = z).$$

We prove propositions 4.4.2 and 4.4.3 in section 4.5.

From these two propositions, for every $j \in \tau_1$,

$$\mathbb{E}(\Pi_1(n_{0:3})) \geqslant c^2 \alpha \mathbb{P}(4P_0 - 1 \in I_t)\, \mathbb{E}(\Pi_j(n_{0:3})).$$

Assume that c is so large that $c^2\alpha\mathbb{P}(4P_0 - 1 \in I_t) \geqslant \eta$. Then, for every $n_{0:3} \in F_c^{(n)}$, for every $j \in \tau_1$,

$$\mathbb{E}(\Pi_1(n_{0:3})) \geqslant \eta\, \mathbb{E}(\Pi_j(n_{0:3})).$$

This implies that

$$\mathbb{P}\left(A_\eta^1 \,\middle|\, N \in F_c^{(n)}\right) = 1,$$

which yields the theorem.

4.5 Proof of propositions 4.4.2 and 4.4.3

4.5.1 Proof of proposition 4.4.2

The proof is decomposed into two intermediate results, stated as lemmata below and using estimates on auxiliary random variables introduced below.

Notation 4.5.1. *For every $n \geqslant 1$ and $t > 0$, let $\Gamma_t(n) = [0, 1/n] \times [t, t + 1/n]$.*
For every $t > 0$, let $\mu_t = q_0^{q_0} q_1^{q_1} q_2^{q_2} q_3^{q_3} = q_0^{q_0} q_1^{3q_1}$ and U_t denote the random variable

$$U_t = \prod_{i \in \pi} (P_i/q_i)^{q_i}.$$

For every $n_{0:3}$, for every $j \in \tau_1$, let $W_j(n_{0:3})$ denote the random variable

$$W_j(n_{0:3}) = P_0^{\Delta_0(n_{0:3})} P_1^{(\Delta_j - \Delta_0/3)(n_{0:3})} P_2^{(\Delta_1 + \Delta_k - 2\Delta_0/3)(n_{0:3})}, \quad \textit{with} \quad \{j,k\} = \tau_1.$$

One sees that

$$U_t = P_0^{q_0} P_1^{q_1} P_2^{2q_1} / \mu_t, \qquad Q_n(t) = \mathbb{P}\left(\mathfrak{T} \in \Gamma_t(n)\right),$$

and

$$W_j = (P_0/P_2)^{\Delta_0} (P_1/P_2)^{\Delta_j - \Delta_0/3}.$$

Lemma 4.5.2. *(1) For every $n_{0:3} \in F_c^{(n)}$, for every $j \in \tau_1$, $W_j(n_{0:3}) \leqslant 1$.*
(2) For every $n_{0:3} \in F_c^{(n)}$, for every $j \in \tau_1$, $W_j(n_{0:3}) \geqslant (q_1)^c$ on the event $\{\mathfrak{T} \in \Gamma_t(n)\}$.
(3) There exists a finite constant κ such that $U_t^n \geqslant \mathrm{e}^{-\kappa}$ uniformly on $n \geqslant 1$ and $\{\mathfrak{T} \in \Gamma_t(n)\}$.

Proof of lemma 4.5.2. (1) For every $\mathfrak{T}$, $P_0 \geqslant P_1 \geqslant P_2$. On $F_c^{(n)}$, $\Delta_0 \leqslant 0$ and for every $j \in \tau_1$, $\Delta_j - \Delta_0/3 \leqslant 0$ hence

$$(P_0/P_1)^{\Delta_0} \leqslant 1, \quad (P_0/P_2)^{\Delta_j - \Delta_0/3} \leqslant 1.$$

This proves the claim.

(2) One has $P_0 \leqslant 1$ everywhere and $P_1 \geqslant q_1$ and $P_2 \geqslant q_1$ on the event $\{\mathfrak{T} \in \Gamma_t(n)\}$. On $F_c^{(n)}$, $\Delta_0 \leqslant 0$ and for every $j \in \tau_1$, $\Delta_j - \Delta_0/3 \leqslant 0$ hence $W_j \geqslant q_2^{-\Delta_j - 2\Delta_0/3}$. Finally, on $F_c^{(n)}$, $\Delta_j + 2\Delta_0/3 \leqslant -c$. This proves the claim.

(3) For every $\mathfrak{T} \in \Gamma_t(n)$, $T_{\mathrm{i}} \geqslant 0$ and $T_{\mathrm{e}} \geqslant t$, hence $P_1 \geqslant q_1$ and $P_2 \geqslant q_2 = q_1$. Likewise, $T_{\mathrm{i}} \leqslant 1/n$ and $T_{\mathrm{e}} \leqslant t + 1/n$ hence $P_0 \geqslant p_0(1/n, t+1/n) \geqslant q_0 - 5\mathrm{e}^{-4t}(1 - \mathrm{e}^{-4/n})/4$. This yields that, for every $n \geqslant 1$ and $\mathfrak{T} \in \Gamma_t(n)$,

$$U_t^n \geqslant (1 - 5\mathrm{e}^{-4t}/(q_0 n))^n \to \exp(-5\mathrm{e}^{-4t}/q_0) > 0,$$

which implies the desired lower bound. □

Lemma 4.5.3. *For every $n_{0:3} \in F_c^{(n)}$, for every $j \in \tau_1$,*

$$\mathbb{E}(\Pi_j(n_{0:3}) \,|\, 4P_0 - 1 \in I_t) \geqslant \mu_t^n Q_n(t) \mathrm{e}^{-O(\sqrt{n})},$$

and

$$\mathbb{E}(\Pi_j(n_{0:3}) \,|\, 4P_0 - 1 \notin I_t) \leqslant \mu_t^n \mathrm{e}^{-(\ell_t^2/32)n}.$$

Proof of lemma 4.5.3. Since $P_0 = p_0(\mathfrak{T})$, for every $\mathfrak{T} \in \Gamma_t(n)$, when n is large, $4P_0 - 1 \in I_t$. Consequently,

$$\mathbb{E}(\Pi_j(n_{0:3}) \,|\, 4P_1 - 1 \in I_t) \geqslant Q_n(t) \mathbb{E}\left(\Pi_j(n_{0:3}) \,|\, \mathfrak{T} \in \Gamma_t(n)\right).$$

On the event $\{\mathfrak{T} \in \Gamma_t(n)\}$,

$$\Pi_j(n_{0:3}) = \mu_t^n U_t^n W_j(n_{0:3})^{\sqrt{n}} \geqslant \mu_t^n \mathrm{e}^{-\kappa} (q_1)^{c\sqrt{n}},$$

from parts (2) and (3) of lemma 4.5.2, which proves the first part of the lemma.

Turning to the second part, let d_{KL} denote the Kullback-Leibler distance between probability measures. When $4P_0 - 1 \notin I_t$,

$$\mathrm{d}_{KL}(q_{0:3}, P_{0:3}) \geqslant (1/2)\|q_{0:3} - P_{0:3}\|_1^2 \geqslant (1/2)(q_0 - P_0)^2 \geqslant \ell_t^2/(32).$$

Note that

$$\Pi_j(n_{0:3}) = \mu_t^n W_j(n_{0:3})^{\sqrt{n}} \mathrm{e}^{-n\mathrm{d}_{KL}(q_{0:3}, P_{0:3})},$$

hence the estimate on $\mathrm{d}_{KL}(q_{0:3}, P_{0:3})$, and part (1) of lemma 4.5.2, imply the second part of the lemma. □

Turning finally to the proof of proposition 4.4.2, we note that $Q_n(t) = \mathrm{e}^{o(n)}$ because the distribution of $\mathfrak{T}$ is tempered. Furthermore, lemma 4.5.3 shows that, when n is large enough,

$$\mathbb{E}(\Pi_j(n_{0:3}) \,|\, 4P_0 - 1 \in I_t) \geqslant \mathbb{E}(\Pi_j(n_{0:3}) \,|\, 4P_0 - 1 \notin I_t).$$

This concludes the proof of proposition 4.4.2.

4.5.2 Proof of proposition 4.4.3

Our proof of proposition 4.4.3 is based on lemma 4.5.5 and proposition 4.5.6 below.

Notation 4.5.4. *For every u in $[0,1]$, let $\zeta(u) = (1+2u)(1-u)^2$. Let U and V denote the random variables defined as*

$$U = (P_1 - P_2)/(1-P_0), \qquad V = \zeta(U).$$

Lemma 4.5.5. *For every $n_{0:3} \in F_c^{(n)}$, for every $j \in \tau_1$,*

$$\frac{\mathbb{E}(\Pi_1(n_{0:3}) \,|\, P_0)}{\mathbb{E}(\Pi_j(n_{0:3}) \,|\, P_0)} \geqslant 4c^2 n \frac{\mathbb{E}(V^s(1-V) \,|\, P_0)}{\mathbb{E}(V^s \,|\, P_0)}, \quad s = (n-n_0)/3.$$

Proof of lemma 4.5.5. Recall that, for every $c > 1$, $F_c^{(n)}$ is

$$F_c^{(n)} = \{n_{0:3} : |n_{0:3}| = n,\ -2c \leqslant \Delta_2(n_{0:3}) \leqslant c, -2c \leqslant \Delta_3(n_{0:3}) \leqslant c, -c \leqslant \Delta_0(n_{0:3}) \leqslant 0\}.$$

Using the Δ variables, one can rewrite Π_1, Π_2 and Π_3 as

$$\Pi_i(n_{0:3}) = P_0^{n_0}(P_1P_2^2)^s (P_1/P_2)^{\Delta_i(n_{0:3})\sqrt{n}}, \quad i = 1,2,3,\ s = (n-n_0)/3.$$

Assume that $n_{0:3} \in F_c^{(n)}$. Then, $\Delta_1(n_{0:3}) \geqslant 2c$, for every $j \in \tau_1$, $\Delta_j(n_{0:3}) \leqslant 0$ and $P_1 \geqslant P_2$, hence

$$\Pi_1(n_{0:3}) \geqslant P_0^{n_0}(P_1P_2^2)^s (P_1/P_2)^{2c\sqrt{n}}, \qquad \Pi_j(n_{0:3}) \leqslant P_0^{n_0}(P_1P_2^2)^s.$$

Furthermore,

$$P_1P_2^2 = (1/27)V(1-P_0)^3, \quad P_1/P_2 = (1+2U)/(1-U),$$

hence for every $j \in \tau_1$,

$$\frac{\mathbb{E}(\Pi_1(n_{0:3}) \,|\, P_0)}{\mathbb{E}(\Pi_j(n_{0:3}) \,|\, P_0)} \geqslant \frac{\mathbb{E}\left(V^s((1+2U)/(1-U))^{2c\sqrt{n}} \,|\, P_0\right)}{\mathbb{E}(V^s \,|\, P_0)}.$$

Direct computations (or lemma 3.2 in Steel and Matsen [SM07]) show that, for every u in $[0,1)$ and every $m \geqslant 3$,

$$((1+2u)/(1-u))^m \geqslant m^2(1-\zeta(u)),$$

hence

$$((1+2U)/(1-U))^{2c\sqrt{n}} \geqslant 4c^2 n(1-V).$$

The conclusion of lemma 4.5.5 follows. □

Proposition 4.5.6. *Assume that the distribution of $\mathfrak{T}$ is tempered. There exists θ and α, both positive and independent of c, such that for every $s \geqslant \theta$, on the event $\{4P_0 - 1 \in I_t\}$,*

$$4s\,\mathbb{E}(V^s(1-V)\,|\,P_0) \geqslant \alpha\,\mathbb{E}(V^s\,|\,P_0).$$

Proof of proposition 4.5.6. We recall that U and V denote the random variables defined as

$$U = (P_1 - P_2)/(1-P_0), \quad V = \zeta(U), \quad \zeta(u) = (1+2u)(1-u)^2.$$

To use proposition 4.3.2, one must compute a Taylor expansion at $v = 1^-$ or, equivalently, at $u = 0^+$, of the conditional probability

$$\mathbb{P}(V \geqslant v\,|\,P_0) = \mathbb{P}(U \leqslant u\,|\,P_0),$$

where $u = \zeta^{-1}(v)$. Besides, for v close to 1,

$$u = \zeta^{-1}(v) = w/\sqrt{3} + w^2/9 + 5w^3/54\sqrt{3} + O(w^4), \quad \text{with } w = \sqrt{1-v}.$$

Since $U = (P_1 - P_2)/(1-P_0)$,

$$\mathbb{P}(U \leqslant u\,|\,4P_0 - 1 = z) = \mathbb{P}\left(S_{\mathrm{e}}(3 - S_{\mathrm{i}}) \leqslant 2s\,|\,S_{\mathrm{e}}S_{\mathrm{i}} = z\right),$$

where we recall that

$$S_{\mathrm{e}} = \mathrm{e}^{-4T_{\mathrm{e}}}, \quad S_{\mathrm{i}} = 1 + 2\mathrm{e}^{-4T_{\mathrm{i}}}, \quad 2s = u(3-z).$$

Keeping the notation given in definition 4.2.4, one has

$$G_s(z) = \mathbb{P}\left(S_{\mathrm{e}}(3 - S_{\mathrm{i}}) \leqslant 2s\,|\,S_{\mathrm{e}}S_{\mathrm{i}} = z\right).$$

Since the distribution of $\mathfrak{T}$ is tempered, there exists n bounded functions $(F_i)_{i=0}^{n-1}$ on I_t, a positive number α, $n+1$ real numbers

$$0 = \varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_{n-1} \leqslant 2 < \varepsilon_n,$$

and two positive numbers κ and s_0 such that for every $0 \leqslant s \leqslant s_0$ and every $z \in I_t$,

$$\left| G_s(z) - \sum_{i=0}^{n-1} F_i(z) s^{\alpha+\varepsilon_i} \right| \leqslant \kappa s^{\alpha+\varepsilon_n}.$$

Combining this with the relation $2s = u(3-z)$ and the expansion of $u = \zeta^{-1}(v)$ along the powers of w, one sees that there exists bounded functions $(f_i)_{i=0}^{n-1}$ on I_t, a positive number κ' and $0 \leqslant v_0 < 1$ such that for every $v_0 \leqslant v \leqslant 1$ and every $z \in I_t$,

$$\left| \mathbb{P}(V \geqslant v\,|\,4P_0 - 1 = z) - \sum_{i=0}^{n-1} f_i(z)(1-v)^{\alpha/2+\varepsilon_i/2} \right| \leqslant \kappa'(1-v)^{\alpha/2+\varepsilon_n/2}.$$

Since the functions f_i are bounded and positive on I_t, proposition 4.3.2 in section 4.3 implies that there exists a positive number θ such that for every $z \in I_t$ and every $s \geqslant \theta$, the conclusion of proposition 4.5.6 holds. □

Assuming this, the proof of proposition 4.4.3 is as follows. Let s, θ and α as in lemma 4.5.5 and proposition 4.5.6. Since $n - n_0 = (1-q_0)n - \Delta_0\sqrt{n} \geqslant (1-q_0)n$ for every $n_{0:3} \in F_c^{(n)}$, $s = (n-n_0)/3 \geqslant \theta$ when n is large enough. Furthermore, $s \leqslant n/3$. Finally, for every $n_{0:3} \in F_c^{(n)}$ with n large enough, on $\{4P_0 - 1 \in I_t\}$, for every $j \in \tau_1$,

$$\mathbb{E}(\Pi_1(n_{0:3}) \,|\, P_0) \geqslant 3c^2\alpha\, \mathbb{E}(\Pi_j(n_{0:3}) \,|\, P_0).$$

This concludes the proof of proposition 4.4.3.

4.6 Proof of propositions 4.2.6, 4.2.8, and 4.2.9

Notation 4.6.1. *Introduce the random variables*

$$(S_\mathrm{e}, S_\mathrm{i}) = \varsigma(T_\mathrm{e}, T_\mathrm{i}), \qquad \varsigma(t_\mathrm{e}, t_\mathrm{i}) = (\mathrm{e}^{-4t_\mathrm{e}}, 1 + 2\mathrm{e}^{-4t_\mathrm{i}}),$$

that is,

$$S_\mathrm{e} = \mathrm{e}^{-4T_\mathrm{e}}, \qquad S_\mathrm{i} = 1 + 2\mathrm{e}^{-4T_\mathrm{i}}.$$

Hence, $G_s(z)$ is also

$$G_s(z) = \mathbb{P}\left(3S_\mathrm{e} \leqslant 2s + z \,|\, S_\mathrm{e}S_\mathrm{i} = z\right).$$

4.6.1 Proof of proposition 4.2.6

The distribution of $(S_\mathrm{e}, S_\mathrm{i})$ has a smooth joint probability density ϖ, defined on $0 < x \leqslant 1 < y \leqslant 3$ by

$$\varpi(x,y) = \frac{\omega \circ \varsigma^{-1}(x,y)}{16x(y-1)}.$$

For tame priors, the probability $Q_n(t)$ introduced in condition (2) of definition 4.2.4 is of order $1/n^2$. Thus condition (2) of definition 4.2.4 holds.

The definition of $G_s(z)$ as a conditional expectation can be rewritten as

$$G_s(z) = \mathbb{P}\left(3S_\mathrm{e} \leqslant 2s + S_\mathrm{e}S_\mathrm{i} \,|\, S_\mathrm{e}S_\mathrm{i} = z\right).$$

Hence, for every measurable bounded function H,

$$\mathbb{E}\left(H(S_\mathrm{e}S_\mathrm{i})\,;\, 3S_\mathrm{e} \leqslant 2s + S_\mathrm{e}S_\mathrm{i}\right) = \mathbb{E}\left(H(S_\mathrm{e}S_\mathrm{i})G_s(S_\mathrm{e}S_\mathrm{i})\right),$$

that is,

$$\iint H(xy)\mathbf{1}\{3x \leqslant 2s + xy\}\varpi(x,y)\mathrm{d}x\mathrm{d}y = \iint H(xy)G_s(xy)\varpi(x,y)\mathrm{d}x\mathrm{d}y.$$

The change of variable $z = xy$ yields

$$\iint H(z)\mathbf{1}\{3x \leqslant 2s+z\}\varpi(x,z/x)\,\mathrm{d}z\mathrm{d}x/x = \iint H(z)G_s(z)\varpi(x,z/x)\,\mathrm{d}z\mathrm{d}x/x.$$

This should hold for every measurable bounded function H, hence one can choose

$$G_s(z) = H_z(s)/H_z(\infty),$$

with

$$H_z(s) = \int \mathbf{1}\{3x \leqslant 2s+z\}\varpi(x,z/x)\,\mathrm{d}x/x.$$

Since $0 \leqslant S_\mathrm{e} \leqslant 1 \leqslant S_\mathrm{i} \leqslant 3$ almost surely, the integral defining $H_z(s)$ may be further restricted to the range $0 \leqslant x \leqslant 1$ and $z/3 \leqslant x \leqslant z$. Finally, for every $s \geqslant 0$ and $z \in [0,3]$,

$$G_s(z) = H_z(s)/H_z(1),$$

where

$$H_z(s) = \int_{m(0,z)}^{m(s,z)} \varpi(x,z/x)\,\mathrm{d}x/x, \quad \text{with } m(s,z) = \min\{1,z,(2s+z)/3\}.$$

Hence, $m(0,z) = z/3$ and, for small positive values of s, $m(s,z) = m(0,z) + 2s/3$. When $0 \leqslant z \leqslant 1$, $m(s,z) \to m(\infty,z) = z$ when $s \to \infty$ and this limit is reached for $s = z$. When $1 \leqslant z \leqslant 3$, $m(s,z) \to m(\infty,z) = 1$ when $s \to \infty$ and this limit is reached for $s = (3-z)/2$. In both cases, $m(\infty,z) = m(1,z)$ hence $H_z(\infty) = H_z(1)$.

Because ω and ς^{-1} are smooth, Taylor-Lagrange formula shows that, for every $s \geqslant 0$ and every fixed z,

$$H_z(s) = H_z(0) + H_z'(0)s + \frac{1}{2}H_z''(0)s^2 + \frac{1}{6}H_z^{(3)}(0)s^3 + \frac{1}{24}\int_0^s (x-s)^3 H_z^{(4)}(s)\mathrm{d}x.$$

Simple computations yield $H_z(0) = 0$ and the values of $H_z'(0)$, $H_z''(0)$ and $H_z^{(3)}(0)$ as combinations of ω and of partial derivatives of ω, evaluated at the point $(\vartheta,0)$, where $3\mathrm{e}^{-4\vartheta} = z$.

Furthermore, the hypothesis on ω ensures that $H_z^{(4)}$ is bounded, in the following sense: there exists positive numbers s_0 and κ_0 such that for every $s \in [0,s_0]$ and every $z \in I_t$,

$$\left|H_z^{(4)}(s)\right| \leqslant 24\kappa_0.$$

Hence, $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ fulfills the first condition to be tempered, with

$$k = 3, \quad \alpha = 1, \quad \varepsilon_1 = 1, \quad \varepsilon_2 = 2, \quad \varepsilon_3 = 3, \quad \kappa = \kappa_0,$$

and, for every $0 \leqslant i \leqslant 2$,

$$F_i(z) = H_z^{(i+1)}(0)/H_z(1).$$

Finally, since ω is smooth, the functions F_i are bounded on I_t.

4.6.2 Proof of proposition 4.2.8

Recall that, using the random variables $S_{\mathrm{e}} = \mathrm{e}^{-4T_{\mathrm{e}}}$ and $S_{\mathrm{i}} = 1 + 2\mathrm{e}^{-4T_{\mathrm{i}}}$, the function G_s is characterized by the fact that, for every measurable bounded function H,

$$\mathbb{E}\left(H(S_{\mathrm{e}}S_{\mathrm{i}}) : S_{\mathrm{e}}(3 - S_{\mathrm{i}}) \leqslant 2s\right) = \mathbb{E}\left(H(S_{\mathrm{e}}S_{\mathrm{i}})G_s(S_{\mathrm{e}}S_{\mathrm{i}})\right).$$

Here, S_{e} and S_{i} are independent, the distribution of S_{e} is uniform on $[0,1]$ and the distribution of S_{i} is discrete with

$$\mathbb{P}(S_{\mathrm{i}} = y_n) = r_n/r.$$

Thus,

$$\sum_n r_n \int_0^1 H(xy_n)\,\mathbf{1}\{x(3 - y_n) \leqslant 2s\}\,\mathrm{d}x = \sum_n r_n \int_0^1 H(xy_n)\,G_s(xy_n)\,\mathrm{d}x.$$

The changes of variable $z = y_n x$ in each integral yield

$$\sum_n (r_n/y_n) \int H(z)\,\mathbf{1}\{z \leqslant y_n\}\mathbf{1}\{3z \leqslant (2s+z)y_n\}\,\mathrm{d}z = \sum_n (r_n/y_n) \int H(z)\,\mathbf{1}\{z \leqslant y_n\}\,G_s(z)\,\mathrm{d}z.$$

This should hold for every measurable bounded function H, hence

$$G_s(z) = H_z(s)/H_z(\infty), \quad H_z(s) = \sum_n (r_n/y_n)\,\mathbf{1}\{z \leqslant y_n\}\mathbf{1}\{3z \leqslant (2s+z)y_n\}.$$

Since $r_n/y_n = n^{-b} - (n+1)^{-b}$ for $n \geqslant 1$, $H_z(s) = n(z,s)^{-b}$ where

$$n(z,s) = \inf\{n \geqslant 1 \,|\, z \leqslant y_n,\, 3z \leqslant (2s+z)y_n\}.$$

Since $y_n \to 3$ when $n \to \infty$, $n(z,s)$ is finite for every $z < 3$ and $s > 0$.

For every $z > 0$, when s is large enough, namely $s \geqslant (3-z)/2$, the condition $3z \leqslant (2s+z)y_n$ becomes useless and

$$n(z,s) = \inf\{n \geqslant 1 \,|\, z \leqslant y_n\},$$

hence $n(z,s)$ and $H_z(s)$ are independent of s. If $z \geqslant 1$, this implies that $n(z,s)$ and $H_z(s)$ are independent of $s \geqslant 1$. If $z < 1$ and $s \geqslant 1$, the conditions $z \leqslant y_n$ and $3z \leqslant (2s+z)y_n$ both hold for every $n \geqslant 1$ hence $n(z,s) = 1$ and $H_z(s) = 1$. In both cases, $H_z(\infty) = H_z(1)$.

We are interested in small positive values of s. For every $z < 3$, when s is small enough, namely $s \leqslant (3-z)/2$, the condition $z \leqslant y_n$ becomes useless and

$$n(z,s) = \inf\{n \geqslant 1 \,|\, 3z \leqslant (2s+z)y_n\},$$

When furthermore $s < z$, $n \geqslant n(z,s)$ is equivalent to the condition

$$n^{-a} \leqslant h(s/z), \quad \text{with} \quad h(u) = -\frac{1}{4}\ln\left(1 - \frac{3u}{1+2u}\right),\ 0 \leqslant u < 1.$$

Finally, for every $s < \min\{z,(3-z)/2\}$, $n(z,s)$ is the unique integer such that

$$n(z,s)-1 < h(s/z)^{-1/a} \leqslant n(z,s).$$

This reads as

$$h(u)^{b/a}[1+h(u)^{1/a}]^{-b} < H_z(1)\,G_s(z) \leqslant h(u)^{b/a}, \quad u = s/z.$$

One sees that the function h is analytic and that $h(u) = (3u/4) + o(u)$ when $u \to 0$, hence,

$$h(u)^{b/a} = (3u/4)^{b/a}(1 + a_1 u + a_2 u^2 + a_3 u^3 + o(u^3)),$$

when $u \to 0$, for given coefficients a_1, a_2 and a_3. Likewise, since $1/a > 3$, $h(u)^{1/a} = o(u^3)$ when $u \to 0$. This implies that

$$[1+h(u)^{1/a}]^{-b} = 1 + o(u^3),$$

hence

$$H_z(1)\,G_s(z) = (3u/4)^{b/a}(1 + a_1 u + a_2 u^2 + a_3 u^3 + o(u^3)).$$

This yields the first part of definition 4.2.4, with

$$k = 3, \quad \alpha = b/a, \quad (\varepsilon_1, \varepsilon_2, \varepsilon_3) = (1,2,3),$$

and

$$F_0(z) = (3/4z)^{b/a}/H_z(1), \quad F_1(z) = a_1 F_0(z)/z, \quad F_1(z) = a_2 F_0(z)/z^2.$$

The remaining step is to get rid of the dependencies over z of our upper bounds. For instance, the reasoning above provides as an error term a multiple of

$$u^{\alpha+3}/H_z(1) = s^{\alpha+3}/(z^{\alpha+3}H_z(1)),$$

instead of a constant multiple of $s^{\alpha+3}$. But $\inf I_t > 0$, hence the $1/z^{\alpha+3}$ contribution is uniformly bounded.

As regards $H_z(1)$, we first note that $H_z(1) = 1$ if $z \leqslant 1$. If $z \geqslant 1$, elementary computations show that $H_z(1) \geqslant c$ if and only if $n(z,1) \leqslant c^{-1/b}$ if and only if $\exp(-c^{a/b}) \geqslant (z-1)/2$, which is implied by the fact that $1 - c^{a/b} \geqslant (z-1)/2$, which is equivalent to the upper bound $c^{a/b} \leqslant (3-z)/2$. Since $\sup I_t < 3$, this can be achieved uniformly over $z \in I_t$ and $1/H_z(1)$ is uniformly bounded as well.

Finally, we asked for an expansion valid on $s \leqslant s_0$, for a fixed s_0, and we proved an expansion valid over $s/z \leqslant u_0$, for a fixed u_0. But one can choose $s_0 = u_0 \inf I_t$. This concludes the proof that the conditions in the first part of definition 4.2.4 hold.

We now prove that the second part of definition 4.2.4 holds. Since T_i and T_e are independent, for every positive integer n,

$$Q_n(t) = \mathbb{P}\,(T_\mathrm{i} \leqslant 1/n)\,\mathbb{P}\,(t \leqslant T_\mathrm{e} \leqslant t + 1/n)\,.$$

One has

$$n\mathbb{P}\left(t \leqslant T_{\mathrm{e}} \leqslant t+1/n\right) \to 4\mathrm{e}^{-4t} \quad \text{when} \quad n \to +\infty,$$

and

$$\frac{1}{r(n^{1/a}+1)^b} \leqslant \mathbb{P}(T_{\mathrm{i}} \leqslant 1/n) \leqslant \frac{3}{rn^{b/a}}.$$

Since $Q_n(t)$ is bounded from below by a multiple of $1/n^{1+b/a}$, the second point of definition 4.2.4 holds.

4.6.3 Proof of Proposition 4.2.9

Recall that, using the random variables $S_{\mathrm{e}} = \mathrm{e}^{-4T_{\mathrm{e}}}$ and $S_{\mathrm{i}} = 1+2\mathrm{e}^{-4T_{\mathrm{i}}}$, the function G_s is characterized by the fact that, for every measurable bounded function H,

$$\mathbb{E}\left(H(S_{\mathrm{e}}S_{\mathrm{i}}) : S_{\mathrm{e}}(3-S_{\mathrm{i}}) \leqslant 2s\right) = \mathbb{E}\left(H(S_{\mathrm{e}}S_{\mathrm{i}})G_s(S_{\mathrm{e}}S_{\mathrm{i}})\right).$$

Case (**i**). Here, S_{e} and S_{i} are independent, the distribution of S_{e} is uniform on $[0,1]$ and S_{i} is a continuous random variable with density

$$\frac{1}{4\tau(s_i-1)}\mathbf{1}\{1+2\mathrm{e}^{-4\tau} \leqslant s_i \leqslant 3\}$$

with respect to the Lebesgue measure. Let ϖ denote the joint probability density defined as

$$\varpi(x,y) = \mathbf{1}\{0 \leqslant x \leqslant 1\}\mathbf{1}\{1+2\mathrm{e}^{-4\tau} \leqslant y \leqslant 3\}\frac{1}{4\tau(y-1)}.$$

Thus,

$$\iint H(xy)\mathbf{1}\{3x \leqslant 2s+xy\}\varpi(x,y)\mathrm{d}x\mathrm{d}y = \iint H(xy)G_s(xy)\varpi(x,y)\mathrm{d}x\mathrm{d}y.$$

The change of variable $z = xy$ yields

$$\iint H(z)\mathbf{1}\{3x \leqslant 2s+z\}\varpi\left(x,z/x\right)\mathrm{d}z\mathrm{d}x/x = \iint H(z)G_s(z)\varpi\left(x,z/x\right)\mathrm{d}z\mathrm{d}x/x.$$

This should hold for every measurable bounded function H, one can choose

$$G_s(z) = H_z(s)/H_z(\infty),$$

with

$$H_z(s) = \int \mathbf{1}\{3x \leqslant 2s+z\}\mathbf{1}\{0 \leqslant x \leqslant 1\}\mathbf{1}\{1+2\mathrm{e}^{-4\tau} \leqslant z/x \leqslant 3\}\mathrm{d}x/(z-x).$$

Finally, for every $s \geqslant 0$ and $z \in [0,3]$,

$$G_s(z) = H_z(s)/H_z(1+\mathrm{e}^{-4\tau}),$$

where

$$H_z(s) = \int_{m(0,z)}^{m(s,z)} \frac{\mathrm{d}x}{z-x}, \quad \text{with } m(s,z) = \min\{1, z/(1+2\mathrm{e}^{-4\tau}), (2s+z)/3\}.$$

Hence, $m(0,z) = z/3$ and, for small positive values of s, $m(s,z) = m(0,z) + 2s/3$. When $0 \leqslant z \leqslant 1+2\mathrm{e}^{-4\tau}$, $m(s,z) \to m(\infty,z) = z/(1+2\mathrm{e}^{-4\tau})$ when $s \to \infty$ and this limit is reached for $s = \frac{1+\mathrm{e}^{-4\tau}}{1+2\mathrm{e}^{-4\tau}} z$. When $1+2\mathrm{e}^{-4\tau} \leqslant z \leqslant 3$, $m(s,z) \to m(\infty,z) = 1$ when $s \to \infty$ and this limit is reached for $s = (3-z)/2$. In both cases, $m(\infty,z) = m(1+\mathrm{e}^{-4\tau},z)$ hence $H_z(\infty) = H_z(1+\mathrm{e}^{-4\tau})$.

For every fixed $0 \leqslant z \leqslant 1+2\mathrm{e}^{-4\tau}$ and every $0 \leqslant s \leqslant \dfrac{1+\mathrm{e}^{-4\tau}}{1+2\mathrm{e}^{-4\tau}} z$,

$$H_z(s) = \log\left(\frac{z}{z-s}\right).$$

For every fixed $1+2\mathrm{e}^{-4\tau} \leqslant z \leqslant 3$ and every $0 \leqslant s \leqslant \dfrac{3-z}{2}$,

$$H_z(s) = \log\left(\frac{z}{z-s}\right).$$

Hence, there exists a positive s_0 such that for every $z \in I_t$ and every $s \in [0, s_0]$,

$$H_z(s) = \log\left(\frac{z}{z-s}\right) = \log\left(1 - \frac{s}{z}\right).$$

Such a function has a Taylor expansion around $s = 0$ with uniformly bounded coefficient over $z \in I_t$. Hence, $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ fulfills the first condition to be tempered.

We now prove that the second part of definition 4.2.4 holds. Since T_i and T_e are independent, for every positive integer n,

$$Q_n(t) = \mathbb{P}(T_\mathrm{i} \leqslant 1/n)\, \mathbb{P}(t \leqslant T_\mathrm{e} \leqslant t + 1/n).$$

One has

$$n\mathbb{P}(t \leqslant T_\mathrm{e} \leqslant t + 1/n) \to 4\mathrm{e}^{-4t} \quad \text{when} \quad n \to +\infty,$$

and

$$\mathbb{P}(T_\mathrm{i} \leqslant 1/n) = \frac{1}{\tau n}, \quad \text{when} \quad n \text{ is large enough.}$$

Since $Q_n(t)$ is bounded from below by a multiple of $1/n^2$, the second point of definition 4.2.4 holds.

Case (**ii**). Here, S_e and S_i are independent, the distribution of S_e is uniform on $[0,1]$ and S_i is a continuous random variable with density

$$\frac{\tau}{8(s_i-1)} \left[-\frac{1}{4}\log\left(\frac{s_i-1}{2}\right)\right]^{-1/2} \mathbf{1}\{1+2\mathrm{e}^{-4} \leqslant s_i < 3\}$$

with respect to the Lebesgue measure.

One can choose

$$G_s(z) = H_z(s)/H_z(\infty),$$

where

$$H_z(s) = \int_{m(0,z)}^{m(s,z)} \left[\frac{-1}{8}\log\left(\frac{z-x}{2x}\right)\right]^{-1/2} \frac{\mathrm{d}x}{z-x},$$

with

$$m(s,z) = \min\{1, z/(1+2\mathrm{e}^{-4}), (2s+z)/3\}.$$

Hence, $m(0,z) = z/3$ and, for small positive values of s, $m(s,z) = m(0,z) + 2s/3$. When $0 \leqslant z \leqslant 1+2\mathrm{e}^{-4}$, $m(s,z) \to m(\infty,z) = z/(1+2\mathrm{e}^{-4})$ when $s \to \infty$ and this limit is reached for $s = \frac{1+\mathrm{e}^{-4}}{1+2\mathrm{e}^{-4}}z$. When $1+2\mathrm{e}^{-4} \leqslant z \leqslant 3$, $m(s,z) \to m(\infty,z) = 1$ when $s \to \infty$ and this limit is reached for $s = (3-z)/2$. In both cases, $m(\infty,z) = m(1+\mathrm{e}^{-4},z)$ hence $H_z(\infty) = H_z(1+\mathrm{e}^{-4})$.

Hence, there exists a positive s_0 such that for every $z \in I_t$ and every $s \in [0, s_0]$,

$$H_z(s) = \int_0^s \frac{1}{(z-x)}\left[\frac{-1}{4}\log\left(1-\frac{3x}{2x+z}\right)\right]^{-1/2}\mathrm{d}x.$$

Hence,

$$H_z(s) = \frac{4}{\sqrt{3z}}\sqrt{z} + \frac{5}{(3z)^{3/2}}s^{3/2} + \frac{9\sqrt{3}}{40z^{5/2}}s^{5/2} + O(s^{7/2}),$$

where $O(s^{7/2})$ is uniformly bounded over $z \in I_t$. Hence, $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ fulfills the first condition to be tempered.

We now prove that the second part of definition 4.2.4 holds. Since T_i and T_e are independent, for every positive integer n,

$$Q_n(t) = \mathbb{P}\left(T_\mathrm{i} \leqslant 1/n\right)\mathbb{P}\left(t \leqslant T_\mathrm{e} \leqslant t+1/n\right).$$

One has

$$n\mathbb{P}\left(t \leqslant T_\mathrm{e} \leqslant t+1/n\right) \to 4\mathrm{e}^{-4t} \quad \text{when} \quad n \to +\infty,$$

and

$$\mathbb{P}(T_\mathrm{i} \leqslant 1/n) = \frac{1}{n^\tau}, \quad \text{when} \quad n \text{ is large enough.}$$

Since $Q_n(t)$ is bounded from below by a multiple of $1/n^{2+\tau}$, the second point of definition 4.2.4 holds.

Case (**iii**). Here, S_e and S_i are independent, the distribution of S_e is uniform on $[0,1]$ and S_i is a continuous random variable with density

$$-\frac{1}{4(s_i-1)}\log\left[-\frac{1}{4}\log\left(\frac{s_i-1}{2}\right)\right]\mathbf{1}\{1+2\mathrm{e}^{-4} \leqslant s_i < 3\}$$

with respect to the Lebesgue measure.

One can choose

$$G_s(z) = H_z(s)/H_z(\infty),$$

where

$$H_z(s) = \int_{m(0,z)}^{m(s,z)} \log\left[\frac{-1}{4}\log\left(\frac{z-x}{2x}\right)\right]\frac{\mathrm{d}x}{z-x},$$

with

$$m(s,z) = \min\{1, z/(1+2\mathrm{e}^{-4}), (2s+z)/3\}.$$

Hence, there exists a positive s_0 such that for every $z \in I_t$ and every $s \in [0, s_0]$,

$$H_z(s) = -\int_0^s \frac{1}{(z-x)}\log\left[\frac{-1}{4}\log\left(1-\frac{3x}{2x+z}\right)\right]\mathrm{d}x.$$

The Taylor expansion around zero of $H_z(s)$ begins with

$$\frac{1}{z}\left(1-\log(3/(4z))-\log(s)\right)s.$$

Hence, $\mathfrak{T} = (T_\mathrm{e}, T_\mathrm{i})$ does not fulfill the first condition to be tempered.

4.7 Proof of proposition 4.3.2

For fixed values of α, $(\gamma_i)_{i=0}^n$ and $(\varepsilon_i)_{i=0}^n$, introduce, for every $t > 0$,

$$M_t^{\pm} = \int_0^1 t v^{t-1} F_{\pm}(v)\,\mathrm{d}v, \quad \text{where } F_{\pm}(v) = \sum_{i=0}^{n-1} \gamma_i (1-v)^{\alpha+\varepsilon_i} \pm \gamma_n (1-v)^{\alpha+\varepsilon_n}.$$

Hence,

$$M_t = \int_0^1 t v^{t-1}\mathbb{P}(V \geqslant v)\,\mathrm{d}v = M_t^{\pm} + \int_0^1 t v^{t-1}[\mathbb{P}(V \geqslant v) - F_{\pm}(v)]\,\mathrm{d}v,$$

and

$$M_t^{\pm} = tB(t, \alpha+1)\left(\sum_{i=1}^{n-1}\gamma_i \Lambda(\varepsilon_i, t)P(\varepsilon_i, t) \pm \gamma_n \Lambda(\varepsilon_n, t)P(\varepsilon_n, t)\right),$$

where

$$\Lambda(\varepsilon, t) = \frac{\Gamma(\{t\}+\alpha+1)}{\Gamma(\{t\}+\alpha+\varepsilon+1)}, \quad P(\varepsilon, t) = \prod_{\ell=1}^{[t]+1}\left(1-\frac{\varepsilon}{\alpha+\varepsilon+\{t\}+\ell}\right),$$

and B denotes the beta function

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

From the control of the distribution of V,

$$M_t^- - \gamma v_0^t \leqslant M_t \leqslant M_t^+ + v_0^t \quad \text{where } \gamma = \sum_{i=0}^{n} |\gamma_i|.$$

Combining this with the general expression of $M_t^{\pm}$ given above, one gets

$$\frac{M_{t+1}}{M_t} \leqslant \frac{(t+1)B(t+1,\alpha+1)\chi_+(t+1) + v_0^{t+1}}{tB(t,\alpha+1)\chi_-(t) - \gamma v_0^t},$$

where

$$\chi_{\pm}(t) = \sum_{i=0}^{n-1} \gamma_i \Lambda(\varepsilon_i, t) P(\varepsilon_i, t) \pm \gamma_n \Lambda(\varepsilon_n, t) P(\varepsilon_n, t).$$

Using the fact that

$$\frac{(t+1)B(t+1,\alpha+1)}{tB(t,\alpha+1)} = \frac{t+1}{t+\alpha+1},$$

and that

$$tB(t,\alpha+1)Q_\alpha(t) \geqslant 1, \quad \text{where } Q_\alpha(t) = \frac{(t+\alpha)(t+\alpha-1)\dots(t+\{\alpha\})}{\Gamma(\alpha+1)},$$

one sees that

$$\frac{M_{t+1}}{M_t} \leqslant \frac{t+1}{t+\alpha+1} \frac{\gamma_0 + \chi_+(t+1) + Q_\alpha(t+1)v_0^{t+1}}{\gamma_0 + \chi_-(t) - \gamma Q_\alpha(t) v_0^t}.$$

Furthermore,

$$\frac{\gamma_0 + \chi_+(t+1) + Q_\alpha(t+1)v_0^{t+1}}{\gamma_0 + \chi_-(t) - Q_\alpha(t)v_0^t} = 1 + \frac{\chi_+(t+1) - \chi_-(t) + \kappa(t)v_0^t}{\gamma_0 + \chi_-(t) - \gamma Q_\alpha(t) v_0^t}.$$

where $\kappa(t) = v_0 Q_\alpha(t+1) + \gamma Q_\alpha(t)$ is a polynomial function in t.

From lemma 4.7.1 below, there exists a positive number C which depend on α and $\varepsilon_{0:n}$ only such that

$$\chi_+(t+1) - \chi_-(t) \leqslant [2\gamma_n + \varepsilon_n \gamma] C t^{-\beta}, \quad \chi_-(t) \geqslant -C\gamma t^{-\varepsilon_1}.$$

where

$$\beta = \min\{\varepsilon_n, 1+\varepsilon_1\}, \qquad 1 < \beta \leqslant 2.$$

Combining these estimates on $\chi_+(t+1)$ and $\chi_-(t)$, one sees that there exists finite continuous functions τ_1 and A of $(\gamma_{0:n}, \alpha, \varepsilon_{0:n})$, such that, for every $t \geqslant \tau_1$,

$$R_t \geqslant \alpha/t - A/t^\beta.$$

Choosing $\tau = \max(\tau_1, \tau_2)$ yields proposition 4.3.2 as soon as, for every $t \geqslant \tau_2$ (recall that $\beta > 1$),

$$2At \leqslant \alpha t^\beta.$$

Lemma 4.7.1. *Let $\beta = \min\{\varepsilon_n, 1+\varepsilon_1\}$. There exists a positive number C, which depends on α and $\varepsilon_{0:n}$ only, such that*

$$\chi_+(t+1) - \chi_-(t) \leqslant [2\gamma_n + \varepsilon_n\gamma]Ct^{-\beta}, \quad \chi_-(t) \geqslant -C\gamma t^{-\varepsilon_1}.$$

Proof of lemma 4.7.1. For every real number $t \geqslant 1$ and every $1 \leqslant i \leqslant n$,

$$\mathrm{e}^{-S(\varepsilon_i,t)-T(\varepsilon_i,t)} \leqslant P(\varepsilon_i,t) \leqslant \mathrm{e}^{-S(\varepsilon_i,t)},$$

where

$$S(\varepsilon,t) = \sum_{\ell=1}^{[t]+1} \frac{\varepsilon}{\alpha+\varepsilon+\{t\}+\ell} \quad \text{and} \quad T(\varepsilon,t) = \sum_{\ell=1}^{[t]+1} \frac{\varepsilon^2}{(\alpha+\varepsilon+\{t\}+\ell)^2}.$$

Thus, there exists two positive real numbers C_i^- and C_i^+ such that for every real number $t \geqslant 1$, $C_i^- \leqslant t^{\varepsilon_i}P(\varepsilon_i,t) \leqslant C_i^+$, and one can choose $C_i^+ = (\alpha+\varepsilon_i+3)^{\varepsilon_i}$.
Let $C = \max\{C_i^+ ; 1 \leqslant i \leqslant n\}$. Using the two relations

$$P(\varepsilon_i,t) - P(\varepsilon_i,t+1) = P(\varepsilon_i,t)\frac{\varepsilon_i}{\alpha+\varepsilon_i+t+2},$$

and

$$P(\varepsilon_n,t) + P(\varepsilon_n,t+1) = P(\varepsilon_n,t)\left(2 - \frac{\varepsilon_n}{\alpha+\varepsilon_n+t+2}\right),$$

one sees that

$$\chi_+(t+1) - \chi_-(t) = 2\gamma_n\Lambda(\varepsilon_n,t)P(\varepsilon_n,t) - \sum_{i=1}^{n} \gamma_i\Lambda(\varepsilon_i,t)P(\varepsilon_i,t)\frac{\varepsilon_i}{\alpha+\varepsilon_i+t+2}.$$

For every $1 \leqslant i \leqslant n$, the function $\Lambda(\varepsilon_i,\cdot)$ is positive and bounded by 1. Hence,

$$\begin{aligned}\chi_+(t+1) - \chi_-(t) &\leqslant 2\gamma_n P(\varepsilon_n,t) + \sum_{i=1}^{n} |\gamma_i| P(\varepsilon_i,t)\frac{\varepsilon_i}{\alpha+\varepsilon_i+t+2} \\ &\leqslant C\left(2\gamma_n t^{-\varepsilon_n} + \gamma\varepsilon_n t^{-(1+\varepsilon_1)}\right),\end{aligned}$$

and the first inequality in the statement of the lemma holds. The same kind of estimates yields

$$\chi_-(t) \geqslant -\sum_{i=0}^{n-1} |\gamma_i|\Lambda(\varepsilon_i,t)P(\varepsilon_i,t) - \gamma_n\Lambda(\varepsilon_n,t)P(\varepsilon_n,t),$$

hence the second inequality holds. This concludes the proof of lemma 4.7.1. □

Chapter 5

Further developments

This chapter presents some natural follow ups of the work of this thesis, and a more ambitious project aiming at developing some new models of evolution.

5.1 Follow ups

5.1.1 Monotonicities

In chapter 3, we worked on a class of dependent models [BGP08]. We provided consistent estimators and asymptotic confidence intervals for the evolutive time between two DNA sequences under a specific model of the class, namely the Jukes Cantor model with CpG influence (JC+CpG). The proof of the results is complete for this specific model, excepted for the estimator based on $[A,A]_{\text{obs}}$. However when the 4×4 matrix of substitution rate is more general and belongs to the RN+YpR class, we need to assume some technical properties ensuring that some key functions are monotone. A natural continuation of my work is to prove these monotonicities, in other words, to solve conjecture 3.3.4.

We now present a different proof of the fact that $t \mapsto [C,C](t)$ is indeed a decreasing diffeomorphism for JC+CpG, and we explain a way to extend the method to deal with $t \mapsto [A,A](t)$.

Introduce the constant matrices Q and D defined by

$$Q = \begin{pmatrix} -8 & -2r & 0 & 0 \\ 1 & -(12+2r) & -r & 1 \\ 0 & 4 & -(16+4r) & 0 \\ 0 & -2r & 0 & -8 \end{pmatrix}, \quad D = \begin{pmatrix} 2(C) \\ (CG) \\ 0 \\ 2(C) \end{pmatrix},$$

and the time dependent vector $W(t)$ defined by

$$W(t) = \begin{pmatrix} [C,C](t) \\ [C*,CG](t) \\ [CG,CG](t) \\ [C*,*G](t) \end{pmatrix}.$$

Proposition 5.1.1. *The evolution of $W(t)$ is ruled by the linear differential system*

$$W'(t) = QW(t) + D.$$

We prove that every component of W is a diffeomorphism. However, some components are increasing functions whereas the others are decreasing functions.

Introduce the time-dependent vector $X(t)$

$$X(t) = \begin{pmatrix} -[C,C]'(t) \\ -[C*,CG]'(t) \\ -[CG,CG]'(t) \\ [C*,*G]'(t) \end{pmatrix}.$$

Proposition 5.1.2. *For every positive t, all the components of $X(t)$ are positive.*

Note that proposition 5.1.2 yields that $t \mapsto [C,C](t)$ is a decreasing diffeomorphism.

Proof of proposition 5.1.2. Introduce the constant matrix P,

$$P = \begin{pmatrix} -8 & -2r & 0 & 0 \\ 1 & -(12+2r) & -r & -1 \\ 0 & 4 & -(16+4r) & 0 \\ 0 & 2r & 0 & -8 \end{pmatrix}.$$

From proposition 5.1.1, one sees that the evolution of $X(t)$ is ruled by the linear differential system

$$X'(t) = PX(t), \quad X(0) = x_0,$$

where

$$x_0 = \begin{pmatrix} 6(C)_* + 2r(CG)_* \\ (10+3r)(CG)_* - (C)_* \\ (12+4r)(CG)_* \\ 2(C)_* - (8+2r)(CG)_* \end{pmatrix} = \frac{2(3+r)}{16+5r} \begin{pmatrix} 4 \\ 1 \\ 2 \\ 0 \end{pmatrix}.$$

Introduce the constant vectors c_1, c_2, and c_3

$$c_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix}, \quad c_2 = \begin{pmatrix} 4+2r \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad c_3 = \begin{pmatrix} 4 \\ 1 \\ 2 \\ 0 \end{pmatrix}.$$

Let C denote the closed cone of $\mathbb{R}^4$ defined as $C = c_1\mathbb{R}_+ + c_2\mathbb{R}_+ + c_3\mathbb{R}_+$, where $\mathbb{R}_+ = [0, +\infty)$.

Since every component of the vectors c_1, c_2, and c_3 is non negative, the cone C is a subset of $(\mathbb{R}_+)^4$. As a consequence, if for every positive t, $X(t)$ belongs to the relative interior of C, proposition 5.1.2 follows.

First, note that x_0 belongs to C. Indeed, $x_0 = 2((3+r)/(16+5r))c_3$.

Second, note that C is included in the hyperplane H orthogonal to n, where n is the constant vector

$$n = \begin{pmatrix} -1 \\ 4+2r \\ -r \\ 1 \end{pmatrix}.$$

Since n is an eigenvector of the transposed of P, we deduce that H is invariant by P. Hence, starting from x_0 in H, $X(t)$ belongs to H for every positive t. It remains to prove that $X(t)$ belongs to the relative interior of C for every positive t.

To this end, note that

$$\begin{aligned} Pc_1 &= -8c_1, \\ Pc_2 &= 2rc_1 - 10c_2 + 2c_3, \\ Pc_3 &= 2rc_1 + 6c_2 - (14+4r)c_3. \end{aligned}$$

The coefficients of c_1 and c_2 in the decomposition along the base (c_1, c_2, c_3) of x_0 are nonnegative, hence there exists a positive T such that for every t in $]0, T[$, $X(t)$ belongs to the relative interior of C.

Now, assume by contradiction that $X(t)$ may escape the cone C. As a consequence, there exists a positive t_1 such that for every $0 < t < t_1$, $X(t)$ is in the relative interior of C, and $X(t_1) = x_1$ belongs to one of the relative faces of C, that is, one of the coefficients of the decomposition of $X(t)$ along the base (c_1, c_2, c_3) is positive for every $0 < t < t_1$, and zero at $t = t_1$.

Without loss of generality, assume that this coefficient is the coefficient of c_1. Hence, x_1 is a linear combination of c_2 and c_3 with nonnegative coefficients, and the coefficient of c_1 in the decomposition of Px_1 along the base (c_1, c_2, c_3) is non negative, since Px_1 is a non negative combination of Pc_2 and Pc_3. As a consequence, the coefficient of c_1 in the decomposition of $X(t)$ along the base (c_1, c_2, c_3) is non decreasing in a neighbourhood of t_1. This is a contradiction with the definition of t_1.

This achieves the proof. □

Turning to the case of $t \mapsto [A,A](t)$, we introduce the constant matrices

$$M = \begin{pmatrix} -8 & -2r & 0 \\ 0 & -(12+2r) & 1 \\ 0 & -2r & -8 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -r & 0 \\ 0 & -2r & 0 & 0 \end{pmatrix},$$

and the time-dependent vector

$$V(t) = \begin{pmatrix} -[A,A](t) \\ [*A,CG](t) \\ [*A,C*](t) + [A,G](t) \end{pmatrix}.$$

Proposition 5.1.3. *The evolution of $Y(t) = V'(t)$ is ruled by the linear differential system*

$$Y'(t) = MY(t) + NX(t).$$

Like in the case of $t \mapsto X(t)$, for every positive t, $Y(t)$ belongs to an hyperplane. However, at the moment we are not able to provide a positive cone containing $Y(t)$ for every positive t.

5.1.2 Numerical simulations

A widely used algorithm to construct phylogenetic trees is `PhyML` [GG03]. In this algorithm, it is necessary to compute an initial phylogenetic tree and this step is based on distance-matrix methods, where the distances are computed from independent substitution models. I plan to simulate the evolution of DNA sequences under a non trivial specific model of the neighbour dependent class and to check the influence of the depencencies on the error made by the classical estimators. If this influence is too important and yields consequences on the topology of the tree, then one could consider to build an extension of `PhyML` including neighbour dependent models.

5.1.3 About estimators

In [Fal10], I provided consistent estimators T_C and T_A for the evolutionary time between an ancestral DNA sequence and a present one evolving under JC+CpG. Each of these estimators is associated to an asymptotic confidence interval. Every estimator T_λ defined as a convex combination of T_C and T_A, that is, $T_\lambda = \lambda T_C + (1-\lambda)T_A$, for $\lambda \in [0,1]$, is also a consistent estimator of the evolutive time between DNA sequences. A natural question is the following: with respect to the parameters of JC+CpG, for which λ do we obtain the smallest asymptotic confidence interval? More generally, one can hope to combine these various estimators T_x, and others, in an optimal and statistically well founded construction.

5.1.4 Bayesian approach

The second approach of phylogeny reconstruction I developed during my PhD is Bayesian. More precisely, I studied one of its unfortunate aspect, called the star paradox.

An open question, suggested to me by Mike Steel, is to provide necessary conditions for the star paradox to occur. In other words, what priors would prevent the star paradox to occur?

Another line of research is to extend Susko's results on posterior probabilities to non continuous priors, that is, to compute the limit of posterior probabilities when the priors are more general.

5.2 Models of neighbour-dependent substitution processes with insertion/deletion mechanisms

Substitutions are not the only way to alter DNA sequences. For example, insertions add one or several extra nucleotides to the DNA sequence, and deletions remove one or several nucleotides from the DNA sequences. One could study stochastic models taking into account these four mechanisms: independent evolutions of the sites, influence of the neighbourhood, insertions and deletions.

Since substitutions are not the only way to alter DNA sequences, we want to add mechanisms of deletions and insertions for instance into the simplest RN+YpR model, the Jukes-Cantor model with CpG influence.

We know that there exists a unique Markov process on the integer line with the substitution rates given above, and that this model satisfies equilibrium and structure properties.

Now, we wonder what happens if we add other mechanisms to the process. For instance we would like to authorize a nucleotide to be deleted with a deletion rate α. On the opposite, we also would like to authorize the insertion of a nucleotide in the DNA sequence at rate β.

The first problem is to define mathematically this stochastic process. For instance, models in the RN+YpR class can be constructed on the integer line or on a finite circle. Here, working on a finite circle with a constant length could require deletions and insertions to happen at the same times, to keep the same number of nucleotides. Otherwise, the length of the circle would be variable, and possibly become zero.

Assuming that the model is mathematically well grounded, the next step is to study the existence and uniqueness of an equilibrium for the process. For instance, how does it depend on the deletion and insertion rates, and does the model exhibit a phase transition as α and β vary, similar to the well known phase transition of the

Ising's model?

Assuming that the questions above are solved for such models with influence and insertion/deletion mechanisms, we would like to compute the stationary frequencies of some polynucleotides when the whole system is ergodic and, finally, to compute some consistent estimators of the times of divergence in the ancestor case or, even better, in the homologous case.

Résumé en français

Dans cette partie, je présente mon travail de thèse en français.

Après avoir fourni quelques rappels de génétique, nous présentons dans un premier temps les principes, le lexique et les objectifs de la reconstruction phylogénétique. Dans un deuxième temps, nous détaillons des modèles probabilistes d'évolution de séquences, avec et sans interaction entre les sites, et nous expliquons comment estimer des distances génétiques pour des séquences d'ADN ayant évolué sous un modèle indépendant. Dans un troisième temps, nous prouvons qu'il est possible de fournir des estimateurs consistants pour des temps d'évolution entre séquences d'ADN régies par des modèles avec dépendance entre les sites. Dans un quatrième temps, nous introduisons rapidement les méthodes bayésiennes en reconstruction phylogénétique, nous présentons un de leurs travers, le « Bayesian star paradox » et nous montrons qu'il est possible d'étendre un résultat de Steel et Matsen sur le sujet à des classes moins contraignantes de lois a priori. Nous terminons en présentant quelques prolongements naturels de ce travail de thèse.

Introduction à la génétique et à la phylogénie moléculaire

Afin de mieux comprendre d'où viennent les questions mathématiques de cette thèse, nous introduisons ici quelques notions de génétique et de phylogénie. Le lecteur trouvera dans les ouvrages suivants un moyen d'approfondir les notions effleurées ici [GL00], [SPAP95a], [SPAP95b], [Yan06], [Gas05] and [GS07].

Évolution moléculaire

Dans cette partie, nous évoquons quelques étapes qui ont mené à la découverte du matériel génétique, principalement l'acide désoxyribonucléique (ADN), la naissance de la phylogénie moléculaire due à la compréhension de la structure chimique de l'ADN, ainsi que les différentes mutations qu'une séquence d'ADN peut subir.

Bref historique de l'évolution au niveau moléculaire

Le concept moderne de théorie de l'évolution a été introduit par Darwin[1] au milieu du XIX^{e} siècle dans le célèbre *De l'origine des espèces*. Il a en effet soutenu l'hypothèse que toutes les espèces vivantes ont évolué au cours du temps à partir d'un seul ou quelques ancêtres communs grâce au processus connu sous le nom de sélection naturelle.

Avant lui, quelques scientifiques avaient formulé d'autres hypothèses sur l'évolution des espèces, notamment Lamarck[2] qui pensait que les individus avaient la faculté de s'adapter pendant leur vie et pouvaient transmettre ces caractères acquis à leur descendance. Cette hypothèse allait à l'encontre de celle de Cuvier[3] qui soutenait la théorie du catastrophisme selon laquelle les espèces s'éteignaient à cause de catastrophes, suivies par la formation de nouvelles espèces considérées comme immuables. Cette période riche pour l'histoire des sciences est délicate à retranscrire, et nous préférons ne pas nous étendre dessus par manque de compétence dans le domaine. Nous voulons simplement rendre compte du fait que la théorie de l'évolution n'est pas apparu uniquement avec Darwin, mais que la qualité de son travail fait qu'il a convaincu ses contemporains du bien-fondé de sa théorie, et qu'il est passé à la postérité.

En revanche, bien que Darwin ait réussi à convaincre que les organismes vivants évoluaient, il était incapable d'expliquer comment les variations chez les espèces se transmettaient de génération en génération et quels mécanismes supportaient cette transmission. Mendel[4], qui est à l'origine de ce qui est aujourd'hui appelé les lois de Mendel définissant la manière dont les gènes se transmettent de génération en génération, ignorait également comment les caractères étaient transmis et s'interrogeait sur l'existence d'un matériel qui aurait pu porter cette information. En 1869, Miesher[5] a découvert dans le noyau des cellules une substance riche en phosphate : la nucléine, renommée après acide désoxyribonucléique. Il a émis l'hypothèse que cette substance ait un rôle dans l'hérédité, mais il n'a pas poussé ses recherches dans cette voie. Après lui, les connaissances sur l'ADN n'ont cessé d'augmenter, et en 1952, l'hypothèse que l'ADN était le vecteur de l'information génétique fût validée.

Le fait de pouvoir cibler où était située l'information génétique a permis d'ouvrir le champ de l'évolution moléculaire, et il s'est dégagé trois axes de recherche dans ce domaine. Tout d'abord, la classification du monde vivant et la reconstruction de

[1] Charles Robert Darwin (12 février 1809 – 19 avril 1882).

[2] Jean-Baptiste Pierre Antoine de Monet, Chevalier de la Marck (1eraoût 1744 – 18 décembre 1829), naturaliste français.

[3] Georges Léopold Chrétien Frédéric Dagobert Cuvier (23 août 1769 – 13 mai 1832), naturaliste et zoologiste français.

[4] Gregor Johann Mendel (20 juillet 1822 – 6 janvier 1884), moine et botaniste autrichien, est communément reconnu comme le père fondateur de la génétique.

[5] Johann Friedrich Miescher (13 août 1844 - 26 août 1895), biologiste suisse.

l'histoire des espèces ont pu se faire au niveau moléculaire, et ces méthodes ont progressivement pris le pas sur les méthodes traditionnelles. Ensuite, il était maintenant possible d'étudier les mécanismes du changement à l'intérieur du matériel génétique. Enfin, la question de l'origine de la vie a pu être abordée avec un regard neuf.

Quelques notions de génomique

Le génome est l'ensemble du matériel génétique d'un individu ou d'une espèce codé dans son ADN (à l'exception de certains virus dont le génome est porté par des molécules d'ARN).

L'ADN est composé de deux brins complémentaires se faisant face, et formant une double hélice. Chaque brin est une longue séquence de quatre éléments possible, nommés nucléotides, parmi lesquels on distingue deux purines, l'adénine (A) et la guanine (G), et deux pyrimidines, la thymine (T) et la cytosine (C). Les nucléotides sont enchaînés entre eux grâce à des liaisons impliquant un groupe phosphate, qu'on appelle des liaisons 3'-5' phosphodiester. Cet enchaînement donne une orientation au brin d'ADN dans le sens 5' vers 3'. La complémentarité des brins fait que les brins sont orientés dans des sens opposés. Tout ceci est illustré sur la figure 5.1.

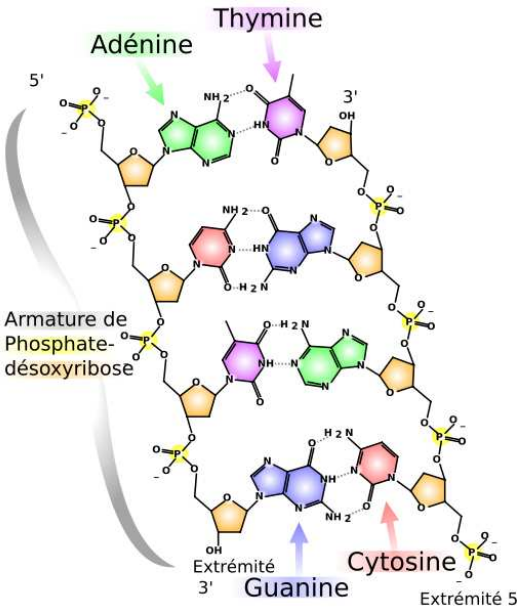


FIG. 5.1 – Structure chimique de l'ADN

Avant chaque division cellulaire, la molécule d'ADN double-brin doit être dupliquée en deux molécules d'ADN filles identiques. Cela assure la transmission de

l'information génétique lors de la reproduction, c'est l'hérédité. Le principe de réplication de l'ADN est illustré sur la figure 5.2.

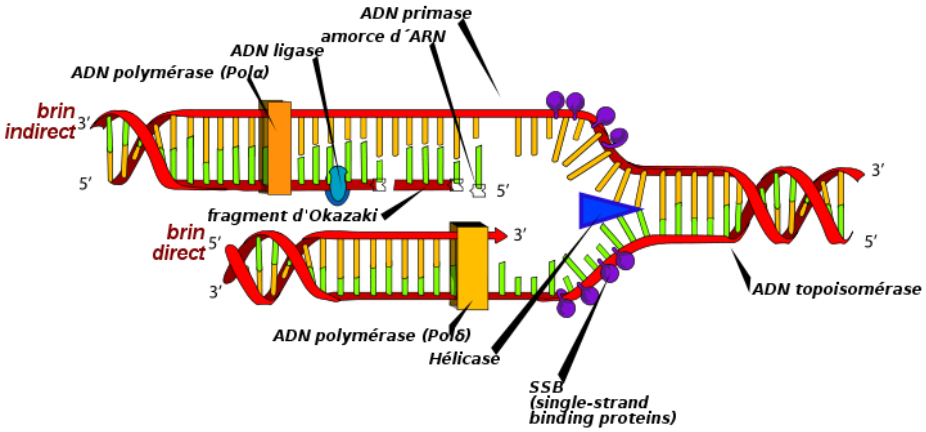


FIG. 5.2 – Schéma de la réplication de la molécule d'ADN

La réplication de l'ADN n'est pas parfaite dans le sens où des erreurs peuvent survenir au cours du processus. Si une erreur se produit dans une cellule germinale, elle est transmise à la descendance, contrairement à une erreur apparue dans une cellule somatique. D'un point de vue évolutif, c'est beaucoup plus intéressant. Les erreurs au cours du processus de réplication ne sont pas les seules causes de mutation de la séquence d'ADN, elles peuvent aussi être provoquées, entre autres, par une exposition à des radiations, par des virus, ou au cours de recombinaisons génétiques. Le type de mutation le plus couramment pris en compte dans les modèles d'évolution est la substitution, c'est à dire le remplacement d'un nucléotide par un autre dans la séquence d'ADN, mais il en existe d'autres, comme les insertions et les délétions par exemple.

La fréquence des mutations est une question épineuse en biologie moléculaire car elle dépend de beaucoup de facteurs. Nous insistons sur un point particulier qui motive cette thèse : les dinucléotides CpG. La notation « CpG » représente ici et dans toute la suite « $5' - CG - 3'$ ». Les cytosines impliquées dans un dinucléotide CpG sont fréquemment méthylées dans le génome des mammifères [Bir80], et cette méthylation entraîne une hausse des fréquences de mutation de CpG vers TpG, et en conséquence, vers CpA sur le brin complémentaire. Ces dinucléotides intéressent particulièrement les biologistes, car il existe des portions dans le génome, appelés îlots CpG, qui possèdent une concentration en CpG plus élevée que dans le reste de la séquence (voir [Bir86], [AB91b], [AB91a]). Il s'avère que ces régions sont souvent des zones fonctionnelles de l'ADN [AB99], et l'existence de telles zones suggère que le phénomène de méthylation est réprimé dans les îlots CpG. La compréhension de ce mécanisme apparaît comme un challenge important en évolution moléculaire.

Phylogénie moléculaire

La phylogénie est l'étude de la formation et de l'évolution des organismes vivants en vue d'établir leur parenté. Auparavant, cette étude était basée sur des critères morphologiques, mais à présent, on utilise les séquences de macromolécules biologiques comme l'ADN ou les protéines. L'idée sous-jacente dans les méthodes traditionnelles, ou bien moléculaires, est la suivante : le degré de ressemblance est corrélé au degré de parenté, c'est à dire que plus deux espèces ou deux séquences d'ADN se ressemblent, plus elles sont proches d'un point de vue évolutif.

On représente couramment une phylogénie, c'est à dire des relations de parenté entre des entités, par un arbre phylogénétique qui est un type particulier de graphe.

Définition. *Un* graphe *est un couple* (V,E) *où* V *est un ensemble d'objets appelés* sommets *ou* noeuds*,* E *est un ensemble d'objets appelés* arêtes *ou* branches*, c'est à dire une liaison entre deux sommets. Un* chemin $(v_0, v_1, \ldots, v_k)$ *est une suite d'éléments de* V *telle que pour chaque entier* i *compris entre* 0 *et* k*, l'arête* (v_i, v_{i+1}) *appartient à* E*. Un* cycle *est un chemin dont les extrémités coïncident. Un* arbre *est un graphe connexe sans cycle.*

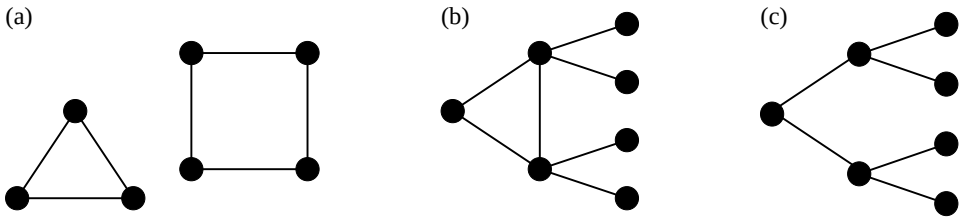


FIG. 5.3 – (a) Graphe non connexe. (b) Graphe connexe avec un cycle. (c) Arbre.

Voici un peu de vocabulaire lié aux arbres et en particulier aux arbres phylogénétiques.

Définition. *Les* feuilles *(noeuds externes) représentent les espèces actuelles, souvent nommés taxa (taxon au singulier), tandis que les noeuds internes représentent les ancêtres éteints pour lesquels les séquences d'ADN ne sont, en général, pas disponibles. L'ancêtre de tous les taxa est appelé la* racine *de l'arbre.*

Nous illustrons ces notions sur la figure suivante.

En phylogénie, on suppose que l'évolution des espèces peut être représentée par un arbre binaire, c'est à dire un arbre où chaque noeud est au plus incident à trois branches. Dans le cas, où un noeud possède plus de trois arêtes incidentes, c'est que nous ne sommes pas capable de lever l'incertitude sur la spéciation. Il faut savoir que l'évolution ne suit pas toujours un arbre, mais nous n'entrerons pas dans ce genre de considération.

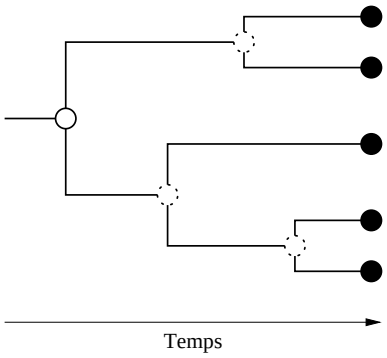


FIG. 5.4 – Un arbre enraciné. Les sommets noirs représentent les feuilles, ceux en pointillés les noeuds internes, et celui en blanc la racine.

Définition. *Un* cladogramme *est un arbre qui représente une topologie, tandis qu'un* phylogramme *est un arbre qui représente une topologie et des longueurs de branches. Un* dendrogramme *est un arbre enraciné où les longueurs des chemins reliant la racine aux espèces sont identiques.*

Les dendrogrammes n'ont un sens en phylogénie que si les espèces évoluent à la même vitesse, et dans ce cas, les longueurs de branches représentent un temps évolutif.

Un des objectifs de la phylogénie est le suivant. En supposant que des espèces actuelles partagent un ancêtre commun et que leur évolution est supportée par un vrai arbre, on cherche à reconstruire un arbre qui approche le mieux possible l'arbre réel, au niveau de la topologie et de la longueur des branches.

Modèles d'évolution de séquences d'ADN

Dans les modèles d'évolution de séquences de nucléotides usuels, on suppose que chaque site de la séquence d'ADN évolue indépendamment des autres sites et suivant un noyau markovien avec des taux de substitution spécifiques. Nous décrivons ici le modèle de Jukes et Cantor et expliquons brièvement comment on estime des distances génétiques pour ce modèle. Nous présentons ensuite une classe de modèles introduite par Bérard, Gouéré et Piau [BGP08] où les sites interagissent entre eux afin de prendre en compte le phénomène de méthylation des dinucléotides CpG dans le génome des mammifères.

Modèle de Jukes et Cantor

Nous présentons ici le modèle de Jukes et Cantor [JC69], introduisons quelques notations et expliquons comment on estime des temps évolutifs pour ce modèle.

Description mathématique

Le modèle de Jukes et Cantor est une chaîne de Markov à temps continu censé modéliser une séquence d'ADN de longueur finie N.

Définition. *L'alphabet des nucléotides est*

$$\mathscr{A} = \{A, T, C, G\}.$$

Ce sont les premières lettres respectives pour Adénine, Thymine, Cytosine et Guanine. Les nucléotides A et G sont des purines représentées par R, les nucléotides T et C des pyrimidines, représentées par Y.

Notons $X_{1:N}(t)$ le vecteur aléatoire $(X_1(t), X_2(t), \ldots, X_N(t))$ à valeurs dans $\mathscr{A}^N$ représentant la valeur des N nucléotides du site 1 au site N et au temps t.

Définition. *Dans le modèle de Jukes et Cantor, le processus $(X_{1:N}(t))_{t\geqslant 0}$ est un processus de Markov dont le générateur infinitésimal Q en chaque site i est donné par la matrice des taux de substitutions*

$$Q = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{array},$$

avec λ un paramètre positif. Chaque site évolue indépendamment des autres.

Proposition. *Pour tout $t \geqslant 0$, la matrice de transition au temps t issue du générateur infinitésimal Q est*

$$P(t) = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} 1-p(t) & p(t)/3 & p(t)/3 & p(t)/3 \\ p(t)/3 & 1-p(t) & p(t)/3 & p(t)/3 \\ p(t)/3 & p(t)/3 & 1-p(t) & p(t)/3 \\ p(t)/3 & p(t)/3 & p(t)/3 & 1-p(t) \end{pmatrix} \end{array}, \quad p(t) = \frac{3}{4}\left(1 - e^{-4\lambda t}\right).$$

Calcul de distance

Notons P_{obs} la quantité observée définie par

$$P_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{X_i(t) \neq X_i(0)\}.$$

Définition. *On note D l'estimateur du temps écoulé défini comme l'unique solution en d de l'équation*

$$P_{\text{obs}} = \frac{3}{4}\left(1 - \mathrm{e}^{-4d}\right).$$

A l'aide de la loi des grands nombres, du théorème limite centrale, de la méthode delta et du lemme de Slustky, on obtient le résultat suivant.

Théorème. *Dans le modèle de Jukes et Cantor, D est un estimateur consistent de* λt *et*

$$(3 - 4P_{\text{obs}})\sqrt{\frac{N}{P_{\text{obs}}(1 - P_{\text{obs}})}}(D - \lambda t) \xrightarrow[N\to+\infty]{d.} \mathcal{N}(0,1).$$

Modèles avec interaction entre les sites

Dans les modèles usuels, il n'y a pas d'interaction entre les sites, chaque nucléotide évolue suivant le même processus de Markov et converge en loi vers la distribution stationnaire associée à la matrice des taux. Proche de l'état d'équilibre, on a alors pour tous nucléotides x et y, $(xy) = (x)(y)$, où (w) désigne la fréquence du mot w dans l'alphabet $\mathscr{A}$. Ceci n'est pas conforme à la réalité. Par exemple, la fréquence du dinucléotide CpG dans le génôme humain est cinq fois plus petite que le produit des fréquences des nucléotides C et G (voir [DG00]).

Les modèles que nous présentons, prennent en compte la nature des sites voisins dans l'évolution d'un site, en particulier nous allons introduire ce que nous appellerons des mutations doubles pour illustrer la méthylation des dinucléotides CpG.

Modèle de Jukes et Cantor avec influence CpG

Le modèle de Jukes et Cantor avec influence CpG est le modèle non trivial le plus simple d'une classe de modèles introduite dans [BGP08], et qui prend en compte cette particularité de l'influence des voisins. Il est construit à l'aide de la superposition de deux mécanismes.

Le premier mécanisme est une évolution indépendante des sites comme dans le modèle de Jukes et Cantor. Le taux de substitution de x par y est de 1 pour tous nucléotides $x \neq y$ dans $\mathscr{A}$.

Un second mécanisme est ajouté, qui décrit les substitutions dues à l'influence du voisinage : on suppose que les taux de substitutions de la cytosine par la thymine et de la guanine par l'adenine sont augmentés de r dans les dinucléotides CpG.

Ceci signifie, par exemple, que tout site occupé par un C dont le voisin de droite n'est pas occupé par un G, change à un taux global de 3, c'est à dire après un temps distribué suivant une loi exponentielle de moyenne $1/3$, comme on peut le voir pour le site situé en position $N-1$ sur la figure 5.5. De plus, le nucléotide C

est remplacé par A, T, ou G avec équiprobabilité. Par contre, un site occupé par un C dont le voisin de droite est occupé par un G, change avec un taux global de $s = 3 + r$, c'est à dire après un temps distribué suivant une loi exponentielle de moyenne $1/s$, comme on peut le voir sur le site situé en position N-3 sur la figure 5.5. Dans cette configuration, le nucleotide C est remplacé par A, T, ou G avec probabilités respectives $1/s$, $(1+r)/s$, et $1/s$.

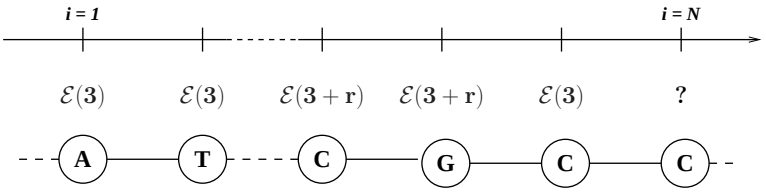


FIG. 5.5 – Un morceau de séquence d'ADN sous le modèle JC+CpG, où l'on représente les horloges exponentielles au-dessus de chaque site.

Modèles RN+YpR

Comme nous le disions précédemment, le modèle de Jukes et Cantor avec influence CpG est en fait issu d'une classe de modèles plus générale introduite dans [BGP08], nommée RN+YpR. Les lettres RN représentent les premières lettres des noms Rzhetsky et Nei, et signifient que la matrice des taux qui régit l'évolution indépendante des sites est de la forme

$$\begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{pmatrix} \cdot & v_T & v_C & w_G \\ v_A & \cdot & w_C & v_G \\ v_A & w_T & \cdot & v_G \\ w_A & v_T & v_C & \cdot \end{pmatrix} \end{array}.$$

Les lettres YpR représentent l'influence du voisinage caractérisée par des taux de substitution supplémentaires pour chaque dinucléotide formé d'une pyrimidine suivi d'une purine, nommément CG, CA, TG, TA, définis ainsi

- Chaque dinucléotide CG est remplacé par CA à taux r_A^C et par TG à taux r_T^G.
- Chaque dinucléotide TA est remplacé par CA à taux r_C^A et par TG à taux r_G^T.
- Chaque dinucléotide CA est remplacé par CG à taux r_G^C et par TA à taux r_T^A.
- Chaque dinucléotide TG est remplacé par CG à taux r_C^G et par TA à taux r_A^T.

Principales propriétés des modèles RN+YpR

Théorème (Bérard, Gouéré and Piau [BGP08])**.** *Pour toute mesure de probabilité ν sur $\mathscr{A}^{\mathbb{Z}}$, il existe un unique processus de Markov $(X(t))_{t\geqslant 0}$ sur $\mathscr{A}^{\mathbb{Z}}$, de loi initiale ν, avec les taux de transition associés ci-dessus.*

Ainsi, pour tout temps t, $X(t)$ décrit toute la séquence et, pour chaque position i, la coordonnée $X_i(t)$ de $X(t)$ est la valeur aléatoire du nucléotide situé en position i et au temps t.

Théorème (Bérard, Gouéré and Piau [BGP08]). *Le processus $(X(t))_{t\geqslant 0}$ est ergodique, son unique mesure stationnaire π sur $\mathscr{A}^{\mathbb{Z}}$ est invariante et ergodique par rapport aux translations de $\mathbb{Z}$, et π alloue une masse strictement positive à chaque mot fini $w=(w_i)_{0\leqslant i\leqslant \ell}$ écrit dans l'alphabet $\mathscr{A}$.*

Les propriétés précédentes viennent d'une représentation de la mesure π.

Théorème (Bérard, Gouéré and Piau [BGP08]). *Il existe une suite i.i.d. $(\xi_i)_{i\in\mathbb{Z}}$ de processus de Poisson marqués, et une application mesurable Ψ à valeurs dans $\mathscr{A}$, tels que si l'on écrit*

$$\Xi_i = \Psi(\xi_{i-1},\xi_i,\xi_{i+1})$$

pour chaque site i dans $\mathbb{Z}$, alors la loi de $(\Xi_i)_{i\in\mathbb{Z}}$ est π.

En particulier, toutes collections $(\Xi_i)_{i\in I}$ et $(\Xi_i)_{i\in J}$ sont independantes dès que les ensembles I et J de $\mathbb{Z}$ sont tels que $|i-j|\geqslant 3$ pour tous sites i dans I et j dans J.

Vers des distances génétiques pour les modèles RN+YpR

Nous considérons le modèle de Jukes et Cantor avec influence CpG (JC+CpG), un modèle de la classe étudiée par Bérard, Gouéré et Piau [BGP08].

Ce modèle est un modèle à temps continu où les séquences d'ADN évoluent sous l'effet combiné de deux mécanismes. Le premier mécanisme est une évolution indépendante des sites comme dans les modèles usuels, c'est à dire ici une matrice 4×4 de taux de substitutions, où chaque substitution se produit au même taux de 1. Un second mécanisme est superposé, qui décrit les taux de substitutions dus à l'influence du voisinage. Nous supposons que les taux de substitutions de CpG vers TpG et vers CpA sont tous les deux augmentés d'un taux supplémentaire noté r.

On note $(x,x)(t)$, pour tout $x\in\{A,C\}$, la fréquence des sites occupés par x au temps 0 dans la séquence ancestrale, et au temps t dans la séquence actuelle, en supposant que les séquences sont alignées. On note $[x,x](t)$,pour tout $x\in\{A,C\}$, la fréquence des sites occupés par x au temps t dans deux séquences actuelles alignées et issues de la même séquence ancestrale.

On note $(x,x)_{\text{obs}}$ et $[x,x]_{\text{obs}}$ les valeurs observées de (x,x) et $[x,x]$ dans deux séquences alignées de longueur N, c'est à dire

$$(x,x)_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} K_i^x(t), \quad \text{avec} \quad K_i^x(t) = \mathbf{1}\{X_i(0)=X_i(t)=x\},$$

et

$$[x,x]_{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} \widetilde{K}_i^x(t), \quad \text{avec} \quad \widetilde{K}_i^x(t) = \mathbf{1}\{X_i^1(t) = X_i^2(t) = x\}.$$

On note T_x et $\widetilde{T}_x$ les estimateurs du temps écoulé t et du temps de divergence t, définis pour chaque $x \in \{A,C\}$ comme les solutions en t des équations

$$(x,x)(t) = (x,x)_{\text{obs}} \quad \text{et} \quad [x,x](t) = [x,x]_{\text{obs}}.$$

Dans cet article, nous prouvons le résultat suivant.

Théorème (Falconnet [Fal10])**.** *Supposons que la séquence ancestrale est en régime stationnaire. Alors, dans le modèle de Jukes et Cantor avec influence CpG, pour tout $x \in \{A,C\}$, quand $N \to +\infty$, il existe une quantité observée explicite α^x_{obs}, respectivement $\widetilde{\alpha}^x_{\text{obs}}$, telle que $\alpha^x_{\text{obs}}\sqrt{N}(T_x - t)$, respectivement $\widetilde{\alpha}^x_{\text{obs}}\sqrt{N}(\widetilde{T}_x - t)$, converge en loi vers la loi normale centrée réduite.*

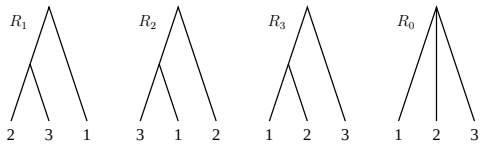
En conséquence de ce théorème, on peut fournir des intervalles de confiance asymptotiques pour le temps écoulé entre une séquence ancestrale et une séquence actuelle, et pour le temps de divergence entre deux séquences actuelles issues d'une même séquence ancestrale.

Un des travers des méthodes bayésiennes : le paradoxe de l'arbre en étoile

En phylogénétique bayésienne, un arbre particulier peut être hautement favorisé alors que les données sont générées à l'aide d'un arbre en étoile, c'est à dire avec une phylogénie non résolue. De récentes études ont mis en lumière le paradoxe dans un contexte très simple, le cas d'un arbre enraciné irrésolu pour trois taxons et deux états, (Yang and Rannala [YR05], Lewis et al. [HLH05]). Kolaczkowski et Thornton [KT06] ont présenté des simulations et suggéré que le support artificiel d'un arbre en particulier pourrait disparaître quand la longueur des séquences devient grande. De précédentes simulations dans[YR05] ont été perturbées par des problèmes numériques, et laissaient inconnue la nature des distributions limites pour les lois a posteriori.

La question statistique derrière le « star paradox » est la suivante : la loi a posteriori de la résolution d'un arbre en étoile devient-elle uniforme quand la longueur des séquences tend vers l'infini, c'est à dire, dans le cas de trois taxons, la loi a posteriori de chaque arbre résolu converge-t-elle vers $1/3$. Dans un récent article, Steel et Matsen ont montré que non, ruinant ainsi l'espoir de Kolaczkowski et Thornton, pour une classe particulière de lois a priori sur la longueur des branches appelée *tame*. Leur résultat est le suivant.

Théorème (Steel & Matsen [SM07])**.** *Considérons des séquences de longueur n générées par un arbre en étoile R_0 sur trois taxons avec une longueur de branche t strictement positive. Soit N le comptage des différents motifs sur les trois taxons. Considérons une loi quelconque sur les trois arbres résolus (R_1, R_2, R_3) dessinés ci dessous et une distribution tame sur les longueurs de leurs branches.*



Alors, pour chaque $i \in \{1,2,3\}$, pour tout ε strictement positif, il existe un δ strictement positif tel que, quand n est assez grand,

$$\mathbb{P}\left((\mathbb{P}(R_i|N) \geqslant 1-\varepsilon\right) \geqslant \delta.$$

Ce résultat a été pris en compte par Yang dans [Yan07] et renforcé par les résultats de Susko dans [Sus08]. Notre principal résultat est que la conclusion de Steel et Matsen se produit pour une classe de lois a priori sur les longueurs de branches plus large, appelée *tempered*, qui peut contenir des accumulations de masses de Dirac.

Théorème (Falconnet [Fal09])**.** *Le résultat de Steel et Matsen a lieu pour toute loi tempere sur les longueurs des branches. Toute loi tame est aussi tempered et la réciproque est fausse.*

Prolongements des résultats de la thèse

Résultat théoriques

Pendant ma thèse, j'ai travaillé sur une classe de modèles avec dépendance [BGP08]. J'ai fourni dans [Fal10] des estimateurs consistants et des intervalles de confiance asymptotiques pour des temps d'évolution entre deux séquences d'ADN pour un modèle spécifique de cette classe : le modèle de Jukes et Cantor avec influence CpG (JC+CpG). La preuve de mes résultats est complète pour ce modèle spécifique. En revanche, dans le cas d'une matrice 4×4 de taux de substitutions plus générale, j'ai eu besoin de supposer que certaines hypothèses techniques étaient vérifiées, notamment la monotonie de certaines fonctions clés. Une poursuite naturelle de mes travaux est de prouver ces hypothèses de monotonie.

Simulations numériques

Actuellement, un algorithme largement utilisé de reconstruction phylogénétique est `PhyML` [GG03]. Dans cet algorithme, il est nécessaire de construire un arbre

phylogénétique initial, et cette étape est réalisée à l'aide de méthodes basées sur des distances entre les séquences, distances calculées à partir de modèles d'évolution indépendants. Mon projet numérique est de simuler l'évolution de séquences d'ADN sous un modèle avec dépendance et de rendre compte de l'influence du voisinage sur les erreurs commises par les estimateurs classiques. Dans la cas où cette influence est importante et a des conséquences sur la topologie de l'arbre, il serait envisageable d'étendre `PhyML` à des modèles avec influence du voisinage.

À propos des estimateurs

Dans [Fal10], j'ai fourni des estimateurs consistants T_C et T_A pour le temps d'évolution entre une séquence d'ADN ancestrale et une séquence actuelle ayant évoluée sous le modèle JC+CpG. Chacun de ces estimateurs est associé à un intervalle de confiance asymptotique. Tout estimateur T_λ défini comme une combinaison convexe de T_C et T_A, c'est à dire $T_\lambda = \lambda T_C + (1-\lambda)T_A$, pour $\lambda \in [0,1]$, est aussi un estimateur consistent du temps d'évolution entre les séquences d'ADN. Une question naturelle est la suivante : en fonction des paramètres du modèle JC+CpG, pour quel λ obtient-on le plus petit intervalle de confiance asymptotique ? Plus généralement, est-il possible de combiner ces différents estimateurs, et d'autres, dans une construction optimale et statistiquement fondée ?

Sur l'approche bayésienne

La seconde approche de la reconstruction phylogénétique que j'ai développée durant ma thèse est une approche bayésienne. Plus précisément, j'ai étudié un de ses aspects négatifs, appelé le "Bayesian star paradox".

Une question ouverte, qui m'a été suggérée par Mike Steel, est de fournir des conditions nécessaires pour que le paradoxe ait lieu. Dit autrement, quelles lois a priori évitent au paradoxe de se produire. Une autre possibilité de recherche est d'étendre les résultats de Susko sur les lois a posteriori au cas de lois a priori possiblement non continues, c'est à dire, de calculer la limite des lois a posteriori quand les lois a priori sont plus générales.

Bibliography

[Aa00] Mark D. Adams and al., *The genome sequence of Drosophila melanogaster*, Science **287** (2000), 2185–2195.

[AB91a] Brahim Aïssani and Giorgio Bernardi, *CpG islands, genes and isochores in the genomes of vertebrates*, Gene **106** (1991), 185–195.

[AB91b] Brahim Aïssani and Giorgio Bernardi, *CpG islands: features and distribution in the genome of vertebrates*, Gene **106** (1991), 173–183.

[AB99] Francisco Antequera and Adrian P. Bird, *CpG islands as genomic footprints of promoters that are associated with replication origins*, Current Biology **9** (1999), R661–R667.

[Ba97] Frederick R. Blattner and al., *The complete genome sequence of Escherichia coli K-12*, Science **277** (1997), 1453–1462.

[BG88] Jean-Pierre Barthélémy and Alain Guénoche, *Les arbres et les représentations de proximité*, Masson, 1988.

[BG10] Jean Bérard and Laurent Guéguen, *Accurate phylogenetic estimation of substitution rates with context-dependent models*, In preparation, 2010.

[BGP08] Jean Bérard, Jean-Baptiste Gouéré, and Didier Piau, *Solvable models of neighbor-dependent nucleotide substitution processes*, Mathematical Biosciences **211** (2008), 56–88.

[Bir80] Adrian P. Bird, *DNA methylation and the frequency of CpG in animal DNA*, Nucleic Acids Research **8** (1980), 1499–1504.

[Bir86] _____, *CpG-rich islands and the function of DNA methylation*, Nature **321** (1986), 209–213.

[CSE67] Luigi L. Cavalli-Sforza and Anthony W. F. Edwards, *Phylogenetic analysis: models and estimation procedures*, Evolution **21** (1967), 550–570.

[DG00] Laurent Duret and Nicolas Galtier, *The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact*, Molecular Biology and Evolution **17** (2000), 1620–1625.

[EW02] Jonathan A. Eisen and Martin Wu, *Phylogenetic analysis and gene functional predictions: phylogenomics in action*, Theoretical Population Biology **61** (2002), 481–487.

[Fa76] Walter Fiers and al., *Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene*, Nature **260** (1976), 500–507.

[Fal09] Mikael Falconnet, *Priors for the Bayesian star paradox*, Available at `http://arxiv.org/abs/0911.0733`, 2009.

[Fal10] _____, *Phylogenetic distances for neighbour dependent substitution processes*, Mathematical Biosciences **224** (2010), 101–108.

[Fel81] Joseph Felsenstein, *Evolutionary trees from DNA sequences : A maximum likelihood approach*, J. Mol. Evol. **17** (1981), 368–376.

[Ga96] André Goffeau and al., *Life with 6000 genes*, Science **274** (1996), 546–567.

[Gas05] Olivier Gascuel, *Mathematics of evolution and phylogeny*, Oxford University Press, 2005.

[GG03] Stéphane Guindon and Olivier Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood*, Systematic Biology **52** (2003), 696–704.

[GGG96] Nicolas Galtier, Manolo Gouy, and Christian Gautier, *Seaview and phylo_win, two graphic tools for sequence alignment and molecular phylogeny*, Comput. Applic. Biosci. **12** (1996), 543–548.

[GL00] Dan Graur and Wen-Hsiung Li, *Fundamentals of Molecular Evolution, Second Edition*, Sinauer Associates, Sunderland, MA, 2000.

[GS07] Olivier Gascuel and Mike Steel, *Reconstructing Evolution - New mathematical and computational methods*, Oxford University Press, 2007.

[HH80] Peter Hall and C. C. Heyde, *Martingale limit theory and its applications*, Academic Press, New York, 1980.

[HKY85] Masahito Hasegawa, Hirohisa Kishino, and T. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA*, J. Mol. Evol. **22** (1985), 160–174.

[HLH05] Mark T. Holder, Paul O. Lewis, and Kent E. Holsinger, *Polytomies and Bayesian phylogenetic inference*, Systematic Biology **54(2)** (2005), 241–253.

[JC69] Thomas Hughes Jukes and Charles R. Cantor, *Mammalian protein metabolism*, ch. Evolution of Protein Molecules, pp. 21–132, Academic Press, New York, 1969.

[JP00] J. L. Jensen and A. Pedersen, *Probabilistic models of DNA sequence evolution with context dependent rates substitution*, Adv. Appl. Prob. **32** (2000), 459–517.

[JTT92] David T. Jones, William R. Taylor, and Janet R. Thornton, *The rapid generation of mutation data matrices from protein sequences*, Comput. Appl. Biosci. **8** (1992), 275–282.

[Ken59] David G. Kendall, *Unitary dilations of one-parameter semigroups of Markov transition operators, and the corresponding integral representations for Markov processes with countable infinity of states*, Proceedings of the London Mathematical Society **9(3)** (1959), 417–431.

[Kim80] Motoo Kimura, *A Simple Method for Estimating Evolutionary Rates of Base Subsitutions Through Comparative Studies of Nucleotide Sequences*, J. Mol. Evol. **10** (1980), 111–120.

[KT06] Bryan Kolaczkowski and Joseph W. Thornton, *Is there a star tree paradox?*, Mol. Biol. Evol. **23** (2006), 1819–1823.

[La01] Eric S. Lander and al., *Human genome*, Nature **409** (2001), 2185–2195.

[Lig85] Thomas M. Liggett, *Interacting Particle System*, Grundlehren der Mathematishen Wissenschaften, vol. 276, Springer-Verlag, New York, 1985.

[NLS+09] Daniel Nätt, Niclas Lindqvist, Henrik Stranneheim, Joakim Lundeberg, Peter A. Torjesen, and Per Jensen, *Inheritance of Acquired Behaviour Adaptations and Brain Gene Expression in Chickens*, PLoS ONE **4(7)** (2009), e6405.

[Nor97] James R. Norris, *Markov Chains*, ch. Continuous-time Markov chains I, Cambridge university press, 1997.

[Ped71] R. A. Pedersen, *DNA content, ribosomal gene multiplicity, and cell size in fish*, Journal of Experimental Zoology **177** (1971), 65–79.

[SK88] James A. Studier and Karl J. Keppler Keppler, *A note on the neighbor-joining algorithm of Saitou and Nei*, Molecular Biology Evolution **5(6)** (1988), 729–731.

[SM07] Mike Steel and Frederick A. Matsen, *The Bayesian 'star paradox' persists for long finite sequences*, Molecular Biology and Evolution **24** (2007), 1075–1079.

[SN87] Naruya Saitou and Masatoshi Nei, *The neighbor-joining method: a new method for reconstructing phylogenetic trees*, Molecular Biology Evolution **4(4)** (1987), 406–425.

[SPAP95a] Michel Solignac, Georges Periquet, Dominique Anxolabéhère, and Claudine Petit, *Génétique et évolution. Tome 1 : La variation, les gènes dans la population*, Hermann, 1995.

[SPAP95b] ———, *Génétique et évolution. Tome 2 : L'espèce, l'évolution moléculaire*, Hermann, 1995.

[SS73] Peter H. A. Sneath and Robert R. Sokal, *Numerical Taxonomy*, W.K. Freeman and Company, San Francisco, 1973.

[Sus08] Edward Susko, *On the distributions of bootstrap support and posterior distributions for a star tree*, Systematic Biology **57(4)** (2008), 602–612.

[Ta06] Gerald A. Tuskan and al., *The genome of Black cottonwood, Populus trichocarpa (Torr. and Gray)*, Science **313** (2006), 1596–1604.

[TN93] Koichiro Tamura and Masatoshi Nei, *Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees*, Molecular Biology and Evolution **10** (1993), 512–526.

[vdV98] Aad W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 1998.

[Yan06] Ziheng Yang, *Computational Molecular Evolution*, Oxford Series in Ecology and Evolution, 2006.

[Yan07] ———, *Fair-Balance paradox, star-tree paradox, and Bayesian phylogenetics*, Molecular Biology and Evolution **24** (2007), 1639–1655.

[YR05] Ziheng Yang and Bruce Rannala, *Branch-length prior influences Bayesian posterior probability of phylogeny*, Syst. Biol. **54(3)** (2005), 455–470.

Abstract. In this thesis, we deal with two problems of phylogeny reconstruction. First, we consider models of nucleotidic substitution processes where the rate of substitution at a given site depends on the state of the neighbours of the site. We estimate the time elapsed between an ancestral sequence at stationarity and a present sequence. Then, assuming that two sequences are issued from a common ancestral sequence at stationarity, we estimate the time since divergence. In the simplest nontrivial case of a Jukes-Cantor model with CpG influence, we provide and justify mathematically consistent estimators in these two settings. We also provide asymptotic confidence intervals, valid for nucleotidic sequences of finite length, and we compute explicit formulas for the estimators and for their confidence intervals. In the general case of an RN model with YpR influence, we extend these results under a proviso, namely that the equation defining the estimator has a unique solution. Second, we show that the Bayesian star paradox, first proved mathematically by Steel and Matsen for a specific class of prior distribution, occurs in a wider context.

Keywords. Markov processes, confidence intervals, DNA sequences, phylogenetic distances, CpG deficiency, phylogenetic trees, Bayesian statistics, star trees.

Résumé. Ce travail de thèse traite de deux problèmes liés aux méthodes de reconstruction d'arbres phylogénétiques. Dans une première partie, nous fournissons des estimateurs consistants ainsi que des intervalles de confiance asymptotiques mathématiquement rigoureux pour le temps d'évolution de séquences d'ADN dans des modèles de substitutions plus réalistes que les modèles usuels, prenant en compte les effets de la méthylation des dinucléotides CpG dans le génome des mammifères. Dans une seconde partie, nous étendons un résultat récent de Steel et Matsen en prouvant qu'un des travers bien connu des méthodes Bayésiennes en phylogénie, appelé « star tree paradox », a en fait lieu dans un cadre plus large que celui de Steel et Matsen.

Mots-clés. Processus de Markov, intervalles de confiance, séquences d'ADN, distances phylogénétiques, déficit en CpG, arbres phylogénétiques, statistiques bayésiennes, arbres en étoile.

MSC Classification. 60J25, 60J28, 62C10, 62F25, 62P10, 92D15, 92D20.