The Statistical Fate of Genomic DNA : Modelling Match Statistics in Different Evolutionary Scenarios

Florian Massip

NNT : 2015SACLS008

THÈSE DE DOCTORAT
DE
L'UNIVERSITÉ PARIS-SACLAY
PRÉPARÉE À
L'UNIVERSITE PARIS-SUD

Unité MaIAGE, INRA de Jouy-en-Josas

ECOLE DOCTORALE N°574
École doctorale de mathématiques Hadamard

Mathématiques aux Interfaces

Par

Mr Florian Massip

The Statistical Fate of Genomic DNA : Modelling Match Statistics in Different Evolutionary Scenarios

Thèse présentée et soutenue à Jouy-en-Josas, le 2 Octobre 2015 :

Composition du Jury :

Mr Amaury Lambert	Professeur au Collège de France	Président
Mr Laurent Duret	Directeur de Recherche, Université Claude Bernard	Rapporteur
Mr Philipp Messer	Professeur, Cornell University	Rapporteur
Mme Christine Dillmann	Professeure, université Paris Sud	Examinateur
Mme Sophie Schbath	Directrice de Recherche, INRA de Jouy-en-josas	Directrice de thèse
Mr Peter Arndt	Directeur de recherche, Max Planck Institut, Berlin	Co-directeur de thèse

PARIS SUD UNIVERSITY

Abstract

The Statistical Fate of Genomic DNA: Modelling Match Statistics in Different Evolutionary Scenarios

by Florian Massip

In this thesis, we study the length distribution of maximal exact matches within and between eukaryotic genomes.These distributions strongly deviate from what one could expect from simple probabilistic models and, surprisingly, present a power-law behavior. To analyze these deviations, we develop mathematical frameworks taking into account complex evolutionary mechanisms and that reproduce the observed deviations. We also implemented *in silico* sequence evolution models that reproduce these behaviors. Finally, we show that we can use our framework to assess the quality of sequences of recently sequenced genomes and to highlight the importance of unexpected biological mechanisms in eukaryotic genomes.

Keywords: Duplications, Scale Free Distributions, Evolutionary Models, Statistical Properties of Genomes

Le but de cette thèse est d'étudier la distribution des tailles des répétitions au sein d'un même génome, ainsi que la distribution des tailles des appariements obtenus en comparant différents génomes. Ces distributions présentent d'importantes déviations par rapport aux prédictions des modèles probabilistes existants. Étonnamment, les déviations observées sont distribuées selon une loi de puissance. Afin d'étudier ce phénomène, nous avons développé des modèles mathématiques prenant en compte des mécanismes évolutifs plus complexes, et qui expliquent les distributions observées. Nous avons aussi implémenté des modèles d'évolution de séquences *in silico* générant des séquences ayant les mêmes propriétés que les génomes étudiés. Enfin, nous avons montré que nos modèles permettent de tester la qualité des génomes récemment séquencés, et de mettre en évidence la prévalence de certains mécanismes évolutifs dans les génomes eucaryotes.

Mots-clefs: Propriétés statistiques des génomes, loi de puissance, duplications, modèles d'évolution

Acknowledgements

I've been told many times — and have witnessed — that working for a PhD can be tough and stressful. I my case, it was quite the opposite, and I know I mainly owe it to my two great supervisors, Sophie and Peter. They both have been available any time I was in need for help, and were really worried about my well being. They also pushed me to give my best and always gave me sound scientific advices. For all this, I am really thankful.

It also wouldn't have been possible to achieve this work without the great help of Misha, his old russian and jewish jokes, and his now famous "haaaa hoooo" exclamations accompanying the development of a new harebrained model.

I am also greatly thankful to all the people at the MPI, for always being so welcoming when I was around. I especially thank Joey, Sina and Tom that made me feel just like home every time I was in Berlin; Anna, Ale, Matt, Brian, Doc C., Mike, Alena, and Prof. Marsico with whom it's always a pleasure to share a beer, as well as Hossein and all the kicker players of the MPI. I also want to thank Yves and Morgane for their help when I first came to Berlin, and for being of great company at the many conferences that we attended together.

I would also like to thank Stéphane Robin and Martin Vingron who gave me the opportunity to work in Berlin in the first place, to meet Peter and Sophie, and start this PhD, as well as Juliette and Martina for their efficiency with all administrative processes.

De ce côté-ci du Rhin, j'aimerais remercier mes collègues de MIG (ainsi que les petits nouveaux de Maiage) de m'avoir supporté pendant ces trois longues années, et particulièrement Cyprien et Emanuele toujours prêts à partir à l'aventure dans les bois josassiens, Sandra, et tous les membres l'équipe Genevol/StatInfOmics pour la diversité de leurs points de vue et la qualité de leurs conseils scientifiques. Merci également à Mahendra pour ses commentaires avisés sur ce manuscrit. Je me dois aussi de remercier tous les apprentis cuisiniers du bâtiment 233, toujours prêts à partager leurs talents, et qui ont été, tout au long de cette thèse, une source d'énergie quasiment inépuisable. Je me dois aussi de remercier la merveilleuse équipe de professeurs de la BCPST du Lycée Fénelon, à qui je dois notamment la découverte des joies de l'interdisciplinarité.

Je n'aurais pas réussi à écrire cette thèse sans le soutien chaleureux de tous mes comparses, et je me dois de remercier pêle-mêle Gabi avec qui j'apprécie toujours autant de briser le mur du çon, Olivier pour m'avoir rappellé quand j'en avais besoin que le monde est vaste, et pour sa 64, Pachou qui sait où chercher des chercheurs qui cherchent, sans avoir à craindre de trouver des chercheurs qui trouvent, Martin pour nos longues discussions électriques au coin du feu et pour son humour corrosif, ainsi que Simon pour son art du contre-pied, ses visites surprises, mais aussi et surtout pour Jean Luc. Bien sûr, je n'oublie pas Adèle, Blanche, Noémie, Toto, Isa, Carole, Damien, Claire, Aude et François qui ont supporté et subi mes humeurs et mon humour, et m'ont soutenu pendant ces trois longues années. Merci à tous.

Un très grand merci plus que mérité également à Sylvain, Papa et Maman, qui se sont préoccupés de mon bien-être au cours de ces 3 années, et qui m'ont apporté un soutien précieux, pas toujours remercié à sa juste valeur.

Enfin, merci à ma petite chérie, qui a su prendre soin de moi et me remonter le moral quand j'en avais besoin, et qui est, en toute objectivité, la plus chouette du monde.

Contents

List of Figures

List of Tables

Abbreviations

MLD	**M**atch **L**ength **D**istribution
CD	**C**overage **D**istribution
WGD	**W**hole **G**enome **D**uplication
bp	**b**ase **p**air
kbp	**K**ilo **b**ase **p**airs
Mbp	**M**ega **b**ase **p**airs
Gbp	**G**iga **b**ase **p**airs
rhs	**R**ight **H**and **S**ide
HMM	**H**idden **M**arkov **M**odel
MLE	**M**aximum **L**ikelihood **E**stimator
TE	**T**ransposable **E**lements
LCNR	**L**ow **C**opy **N**umber **R**epeat
iid	**i**ndependent and **i**dentically **d**istributed
MEM	**M**aximal **E**xact **M**atch

Chapter 1

Introduction

In this Chapter we present twice the same concepts, first in French and then in English. Readers should read one of the two only.

1.1 Version française de l'introduction

Dans cette thèse, nous allons étudier la distribution des tailles des répétitions au sein d'un même génome, ainsi que la distribution des tailles des appariements obtenus en comparant les génomes de différentes espèces. Ces distributions présentent d'importantes déviations par rapport aux prédictions des modèles probabilistes existants et semblent correspondre à une loi de puissance. Nous allons montrer que des modèles évolutifs simples permettent d'expliquer ces déviations. Dans cette introduction, nous commencerons par montrer, par le biais d'exemples historiques, que la mise en évidence et l'explication de certaines propriétés statistiques des génomes ont permis des avancées majeures dans la compréhension des processus génétiques. L'approche scientifique que nous développerons dans cette thèse est héritée de ces exemples historiques. Dans un second temps, nous introduirons certains processus biologiques que nous étudierons plus particulièrement.

1.1.1 Les propriétés statistiques des génomes

Le génome d'un organisme est défini comme la totalité du matériel génétique qu'il possède. Il contient toutes les informations permettant à un individu de se développer et se différencier depuis le stade unicellulaire, et de se reproduire. L'information génétique est transmise d'une génération à la suivante au cours de la reproduction.

Le support de l'information génétique est un long polymère, l'acide désoxyribonucléique (l'ADN), composé d'un enchainement de quatre unités de bases appelées nucléotides. Chaque nucléotide est lui-même composé d'un sucre, d'un groupement phosphate, et d'une base azotée. Les quatre différentes sortes de nucléotides sont toutes composées du même sucre et du même groupement phosphate, mais diffèrent au niveau des bases azotées, pour lesquelles il existe quatre formes chimiques possibles. Les quatre bases azotées possible, l'Adenine (A), la Cytosine (C), la Guanine (G) et la Thymine (T), sont elles-mêmes divisées en deux groupes chimiques, les purines (A et G), qui sont constituées de deux cycles aromatiques, et les pyrimidines (C et T) qui n'en possèdent qu'un, et sont, pour cette raison, plus petites (la forme chimique complète des quatre bases azotées est détaillée en figure 1.1). Les cellules des organismes vivants ont la capacité d'interpréter les informations complexes contenues dans l'ADN afin de produire des molécules d'ARN (au cours d'un procédé appelé la transcription), qui pourront elles-mêmes être traduites par la cellule afin de former des protéines. Ce sont ces dernières qui effectuent réellement la plupart des fonctions biologiques des cellules. Pour cette raison, l'ADN est souvent comparé à un livre de recettes, écrit dans un alphabet de quatre lettres, et dont la langue serait le code génétique (code qui permet la traduction de molécules d'ARN en protéines).

1.1.1.1 La première règle de parité de Chargaff

Dès les début de la génétique, et avant même que la fonction de l'ADN ait été formellement établie, les scientifiques se sont intéressés aux propriétés statistiques

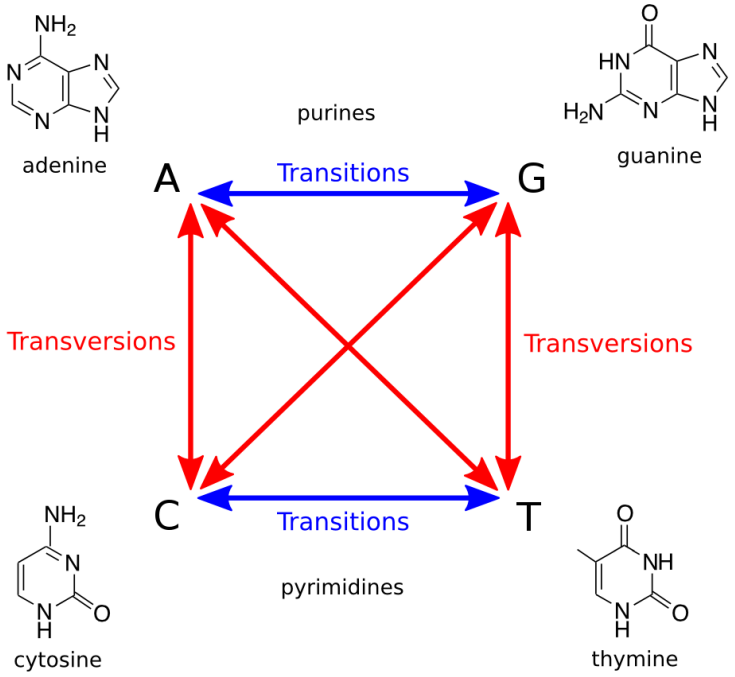

FIGURE 1.1: La classification des mutations en transversions (en rouge) et transitions (en bleu). La structure chimique complète des quatre bases est aussi présentée. Cette image a été emprunté au site internet Rosalind (http://www.rosalind.info).

de l'ADN, dans le but de mieux comprendre son fonctionnement.

Une des premières, et probablement des plus importantes de ces propriétés a été observée par Erwin Chargaff. En étudiant les proportions de chaque nucléotide présent dans les cellules de nombreuses espèces, Chargaff a découvert que ces proportions obéissaient à une règle simple (règle que l'on appelle aujourd'hui la première règle de parité de Chargaff), à savoir que [1]:

$$\begin{cases} \dfrac{n_A}{n_T} = 1 \\ \dfrac{n_C}{n_G} = 1 \end{cases} \tag{1.1}$$

où n_a représente le nombre d'occurrence du nucléotide a. Le tableau 1.1 présente des exemples du nombre de bases obtenu par Chargaff pour différentes espèces. C'est à l'aide de cette observation et des images de cristallographie aux rayons X de l'ADN obtenues par Franklin et Gosling [2] que la structure de l'ADN a été

	% de bases				Ratios		
Espèce	A	G	C	T	A/T	G/C	%GC
ψX174	24.0	23.3	21.5	31.2	0.77	1.08	44.8
Maïs	26.8	22.8	23.2	27.2	0.99	0.98	46.1
Poulpe	33.2	17.6	17.6	31.6	1.05	1.00	35.2
Poulet	28.0	22.0	21.6	28.4	0.99	1.02	43.7
Rat	28.6	21.4	20.5	28.4	1.01	1.00	42.9
Humain	29.3	20.7	20.0	30.0	0.98	1.04	40.7

TABLE 1.1: Ce tableau présente un échantillon des données récoltées par Chargaff en 1952, et représente les compositions dans les quatre nucléotides obtenues pour les génomes de différentes espèces. Le tableau original provient de Bansal [5].

découverte par Watson et Crick [3]. Selon leur modèle, les molécules d'ADN (aussi appelés brins d'ADN) sont associées dans les cellules par paires complémentaires, de manière à ce que chacun des nucléotides d'un brin soit couplé à un nucléotide du brin complémentaire. Les deux molécules d'ADN ainsi associées sont liées l'une à l'autre par des interactions chimiques entre les nucléotides appariés. Du fait de contraintes stériques, les appariements de deux purines ou de deux pyrimidines stabilisent beaucoup moins l'ADN que les appariements entre une purine et une pyrimidine. Il y a donc quatre appariements possibles: A avec T, A avec C, C avec G, et C avec T. Par ailleurs, d'autres interactions chimiques (interactions pi-pi et liaisons hydrogènes [4]) favorisent l'association de A avec T d'une part et de C avec G de l'autre. De ce fait, dans les cellules vivantes, les nucléotides d'adénine sont toujours appariés à des thymines, et les cytosines sont toujours appariées à des guanines. Pour cette raison, dans toutes les cellules vivantes, les proportions en A et en T d'une part ainsi que les proportions en C et en G sont toujours égales, ce qui explique la première règle de parité de Chargaff.

L'appariement des molécules d'ADN par paires joue un rôle central dans un grand nombre de processus biologiques vitaux pour les cellules. De ce fait, la découverte de la structure de l'ADN a permis des avancés significatives dans la

compréhension du fonctionnement des cellules vivantes. Par exemple, à cause des règles d'appariement des nucléotides, les deux brins d'ADN appariés sont complémentaires et contiennent la même information. Ainsi, à partir d'un seul des deux brins, il est possible de totalement reconstruire le brin complémentaire. Cette propriété est utilisée par la cellule pour la réplication de l'ADN. Grâce à des protéines spécifiques, la cellule sépare les deux brins complémentaires, de sorte que chacun des deux brins peut ensuite servir de modèle à la production d'une nouvelle molécule. Une fois les deux nouvelles molécules produites, la cellule contient deux dimères d'ADN et peut alors se diviser pour donner naissance à deux cellules contenant des molécules d'ADN identiques.

Le processus que nous venons de décrire permet à l'ADN de se répliquer avec une très grande fidélité. Toutefois, il arrive que des erreurs se produisent au cours de ce processus. Notamment, au cours de la réplication, une base peut être introduite par erreur à la place d'une autre. Dans ce cas, la nouvelle molécule d'ADN diffère de la molécule parentale au niveau d'une position (ou locus). Comme ce type d'événement, appelé mutation ponctuelle, peut affecter n'importe quelle base, et la remplacer par l'une des trois autres, il y a douze sortes de mutations possibles. Ces mutations sont relativement rares, et ont, la plupart du temps, des effets délétères pour la cellule. Toutefois, sur le long terme, elles sont aussi le moteur de l'évolution des espèces permettant aux organismes de s'adapter aux changements de leur environnement. Par ailleurs, ce type de mutation ne se produit pas uniquement pendant la réplication de l'ADN, et différents processus chimiques peuvent entrainer la transformation d'une base en une autre au cours de la vie de la cellule. Étant donné que les bases azotées appartenant à un même groupe chimique sont similaires entre elles, et assez différentes des deux autres bases, les quatre mutations qui transforment une purine en une autre purine ou une pyrimidine en une autre pyrimidine (ces quatre mutations sont appelées transitions, voir Fig. 1.1) sont plus probables que les 8 autres mutations (que l'on appelle des transversions).

Lorsqu'une telle mutation se produit sur l'un des deux brins, au niveau du nucléotide qui vient d'être inséré, les deux bases azotées ne sont plus correctement appariées, et la structure de l'ADN est modifiée localement. Certaines protéines présentes

dans les cellules sont capables de repérer ces mis-appariements, et de les réparer en échangeant l'un des deux nucléotides. Toutefois, la plupart du temps, la cellule n'est pas capable de distinguer lequel des deux nucléotides a subi une transformation. Ainsi, l'un des deux nucléotides est remplacé au hasard afin que les deux nucléotides soient à nouveau correctement appariés. Ainsi, la moitié seulement des mutations sont réparées, tandis qu'une fois sur deux, la mutation est intégrée.

1.1.1.2 Modéliser les séquences d'ADN

Les génomes des individus d'une même espèces sont très similaires, de sorte qu'il est possible de construire ce que l'on appelle le génome de référence d'une espèce, représentant le génome typique d'un individu. La taille des génomes varie fortement d'une espèce à l'autre, allant de quelques centaines de millier de paires de bases (kbps) pour les plus petits génomes à une centaine de milliard de paires de base (Gbps) pour les plus grands. La taille typique du génome d'un mammifère est de l'ordre de plusieurs Gbps (3.2 Gbps pour le génome humain par exemple), ce qui est comparable, en terme de nombre de lettre, à la taille totale de la version française de Wikipédia (qui selon Wikipédia elle-même, était constituée d'environ 4.2 milliards de lettre en Juin 2015). Toutefois, à la différence de Wikipédia, dont chaque lettre appartient à un mot ayant un sens, seule une petite fraction des génomes des eucaryotes est porteuse de sens. Par exemple, on estime que 1% seulement du génome humain est véritablement fonctionnel (bien que les fonctions potentielles du reste du génome soient l'objet d'un débat passionné au sein de la communauté scientifique [6–8]). Le fait qu'une grande partie des génomes soit non-fonctionnelle permet ainsi d'expliquer pourquoi la taille des génomes varie tant d'une espèce à l'autre, ainsi que la faible corrélation qui a été observée entre la taille du génome d'un organisme et son apparente complexité.

Les génomes sont donc des objets de grande taille, composés d'unités répétées (les nucléotides), ce qui en fait des objets d'étude idéaux pour l'analyse statistique. On peut par exemple chercher à identifier dans les génomes des petits segments (que l'on appellera des mots par la suite) particulièrement rares ou particulièrement

fréquents dans un génome, dans l'idée que ces fréquences exceptionnelles sont la signature de fonctions biologiques [9–11]. Ainsi, on s'attend à ce que certains mots encodant des fonctions particulièrement importantes (comme par exemple les mots qui marquent le point départ de la transcription) soient particulièrement fréquents, tandis que d'autres mots, ayant des conséquences délétères, seraient extrêmement rares, voire totalement absents. Mais pour décider si la fréquence d'un mot est exceptionnelle ou non, il est tout d'abord nécessaire de développer un modèle neutre avec lequel il sera possible de comparer les observations faites dans les génomes réels. Dans les paragraphes qui suivent, nous allons décrire certains des modèles qui ont été développés dans ce but.

Séquences aléatoires — On peut donc représenter une séquence d'ADN $\mathcal{S} = (s_1, ..., s_L)$ par une chaine de L lettres appartenant à l'alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Dans ce cas, le modèle le plus simple (appelé le modèle iid pour indépendant et identiquement distribué) consiste à choisir toutes les lettres s_i indépendamment les unes des autres, en considérant que la probabilité d'apparition f_a d'une base a est la même à chaque position de la séquence. Ce modèle a donc 5 paramètres, qui sont L la longueur de la séquence, ainsi que la fréquence d'apparition de chacune des quatre bases f_A, f_T, f_C et f_G. Dans ce cas, on peut écrire:

$$P(s_i = a) = f_a, \text{ avec } a \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}, \ \forall i \in 1, \ldots, L. \tag{1.2}$$

Un mot de k-lettres, W (aussi appelé un k-mer), est défini comme une sous-séquence de k lettres consécutives. Une séquence de taille L possède $L - k + 1$ mots de longueur k. Si l'on définit $W_i = (s_i, \ldots, s_{i+k-1})$ comme le k-mer dont la première lettre se situe à la position i, la probabilité de trouver un k-mer $W = (w_1, ..., w_k)$ à la position i d'une séquence de longueur L dans le modèle iid est simplement égale à

$$P(W_i = W) = \prod_{j=1}^{k} P(s_{i+j-1} = w_j) = \prod_{j=1}^{k} f_{w_j}. \tag{1.3}$$

En utilisant cette formule, on peut montrer que n_k, défini comme le nombre d'occurrences d'un mot de taille k donné (toujours dans une séquence de longueur L) vaut

$$n_k = (L - k + 1) \prod_{j=1}^{k} f_{w_j}. \tag{1.4}$$

Modèles de Markov — *Cette section s'inspire grandement du livre* ADN, mots et modèles [12] *où l'on trouvera davantage de détails sur la modélisation de séquence d'ADN par des chaines de Markov.*

Afin de répondre aux besoins grandissants de l'analyse statistique des séquences biologiques, des modèles plus sophistiqués ont été développés, et ont permis des avancées majeures dans la compréhension des lois régissant le fonctionnement des génomes. Parmi tous les modèles développés, les chaines de Markov ont été particulièrement utilisées, et seront pour cette raison l'objet de cette section. Les modèles utilisant des chaines de Markov permettent de modéliser un grand nombre de phénomènes, et ont fait l'objet de nombreuses études (voir notamment Karlin and Taylor [13] pour une discussion plus générale à propos de ces modèles).

Les chaines de Markov du premier ordre, qui sont les chaînes de Markov les plus simples, sont des processus dans lesquels l'état de l'élément i dépend uniquement de l'état de l'élément $i-1$, et est indépendant de tous les autres états. En appliquant ce modèle au cas d'une séquence d'ADN, on modélise la séquence comme une chaine de L lettres où la valeur s_i de la lettre présente à la position i dépend uniquement de la valeur de son voisin de gauche, s_{i-1}. Comme il y a quatre valeurs (ou états) possibles pour chaque base s_i, il y a 16 couples (s_{i-1}, s_i) possibles et on définit donc 16 probabilités de transitions $p_{a\to b}$ de passer d'une base $a \in \mathcal{A}$ à une base $b \in \mathcal{A}$ par

$$p_{a\to b} = P(s_i = b | s_{i-1} = a). \tag{1.5}$$

Dans une séquence d'ADN donnée, il est possible d'estimer ces probabilités en utilisant la méthode du maximum de vraisemblance. La vraisemblance d'un modèle est définie comme la probabilité d'observer une série d'événements dans un modèle

donné. Ainsi, la valeur des paramètres qui maximisent la valeur de la vraisemblance donnent des estimateurs de bonne qualité des paramètres, si tant est que le modèle soit correct. Dans le modèle défini précédemment, les estimateurs des probabilités de transition calculés à l'aide du maximum de vraisemblance sont donnés par:

$$p_{a\to b} = \frac{n_{ab}}{\sum\limits_{c\in\mathcal{A}} n_{ac}}, \tag{1.6}$$

où n_{ab} est défini comme le nombre d'occurrences du mot de deux lettres $W = ab$. Ainsi, ce modèle prend comme paramètres les fréquences de tous les di-nucléotides, et est une extension du modèle précédent dans lequel les seuls paramètres étaient les fréquences des mots de une lettre (on notera qu'à partir des fréquences des mots de deux lettres, on peut aisément calculer les fréquences des mots d'une seule lettre). En suivant la même procédure, on peut aussi définir une chaine de Markov d'ordre m, dans laquelle la valeur d'un nucléotide s_i dépend des valeurs des m nucléotides situés en amont. On définit dans ce cas 4^m probabilités de transition, et on peut montrer que les estimateurs du maximum de vraisemblance dépendent des fréquences des mots de taille $m+1$. Ainsi, plus l'ordre de la chaîne de Markov est élévé, plus le nombre de paramètres est élevé et plus la séquence d'ADN est précisément décrite.

Toutefois, la qualité d'un modèle ne se mesure pas toujours à son degré de précision. Ces modèles ont été développés dans le but de calculer la fréquence attendue des mots de tailles k afin de différencier les mots exceptionnels de ceux dont la fréquence est proche de celle qui est attendue. De ce fait, si l'on modélise la séquence d'ADN à l'aide d'une chaîne de Markov d'ordre k, on ne pourra trouver aucun mot de taille k dont la fréquence est exceptionnelle. Une autre manière de se rendre compte qu'un modèle trop précis peut être inutile est de considérer le cas extrême consistant à modéliser une séquence de taille L par une chaîne de Markov d'ordre $L-1$. Dans ce cas, la chaine de Markov reproduit simplement la séquence à l'aide de laquelle les paramètres ont été calculés et sera ainsi totalement inutile.

Chaines de Markov hétérogènes — Dans tous les modèles de Markov que nous avons décrits, les probabilités de transitions sont les mêmes dans toute la séquence (on les appelle pour cette raison des chaînes de Markov homogènes). Toutefois, nous avons vu aussi que seule une petite proportion des génomes eucaryotes est fonctionnelle. Or les régions fonctionnelles et non-fonctionnelles ne sont pas soumises aux mêmes contraintes, et les chaines de Markov homogènes ne permettent donc pas de modéliser cette hétérogénéité des séquences d'ADN. Pour prendre en compte cette propriété, des chaines de Markov hétérogènes, dans lesquelles les probabilités de transition changent d'une région à une autre, ont été développées. Dans le cas de la modélisation des séquences d'ADN, une sous classe de chaines de Markov hétérogènes, les chaines de Markov cachées, sont particulièrement employées, et sont notamment utilisées pour prédire la position des régions codantes des génomes à l'aide de la séquence uniquement [14].

Modéliser l'évolution des séquences d'ADN — Les chaines de Markov sont également souvent employées pour décrire l'évolution d'un processus au cours du temps. Dans le cas de l'évolution au cours du temps d'une molécule d'ADN, on suppose le plus souvent que les différents nucléotides d'une même séquence évoluent indépendamment les uns des autres, et on utilise des chaines de Markov du premier ordre. Cela revient à considérer que le taux de mutation à une position donnée de la séquence dépend du nucléotide observé à l'instant présent, et n'est pas influencé par les états des nucléotides à cette même position dans le passé. Cette hypothèse est réaliste compte tenu du fait que dans les génomes, ces informations (les différents états du nucléotide dans le passé) ne sont pas conservées.

Pour modéliser l'évolution d'une séquence $\mathcal{S}(t) = (s_1(t), ..., s_L(t))$ au cours du temps, on utilise alors une chaîne de Markov en temps continu. Dans un tel modèle, on peut décrire l'évolution du système à l'aide de l'équation maitresse suivante:

$$\frac{\delta}{\delta t}\rho_a(t) = \sum_{b \neq a} \rho_b(t) Q_{ba} - \rho_a(t) \sum_{b \neq a} Q_{ab} \tag{1.7}$$

où $(a, b) \in \mathcal{A}^2$ sont des nucléotides et Q_{ab} est le taux de substitution d'un nucléotide a par un nucléotide b de sorte qu'en un temps infinitésimal δt, la probabilité d'une substitution de a vers b est $Q_{ab}\delta t$.

On peut alors simplifier l'écriture de ces équations à l'aide de l'écriture matricielle et on obtient:

$$\frac{\delta}{\delta t}\rho = \rho(t)Q \tag{1.8}$$

avec $\rho(t)$ un vecteur ligne de dimension 4 tel que:

$$\rho(t) = \Big(\rho_A(t),\, \rho_C(t),\, \rho_G(t),\, \rho_T(t) \Big) \tag{1.9}$$

et Q une matrice carrée de dimension 4 définie par:

$$Q = \begin{pmatrix} \bullet & Q_{CA} & Q_{GA} & Q_{TA} \\ Q_{AC} & \bullet & Q_{GC} & Q_{TC} \\ Q_{AG} & Q_{CG} & \bullet & Q_{TG} \\ Q_{AT} & Q_{CT} & Q_{GT} & \bullet \end{pmatrix} \tag{1.10}$$

où les termes de la diagonale, notés $\bullet$, sont définis de sorte que la somme de chaque colonne soit nulle, c'est à dire:

$$Q_{aa} = -\sum_{a \neq b} Q_{ab}. \tag{1.11}$$

La solution de ces équations différentielles est connue et est telle que:

$$\rho(t) = \rho_0 P(t) \tag{1.12}$$

où

$$P(t) = \exp(Qt) = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!}, \tag{1.13}$$

où $P_{ab}(t)$, la valeur prise par $P(t)$ aux coordonnées a, b, est la probabilité qu'un nucléotide passe de l'état a à l'état b pendant un intervalle de temps fini t, et ρ_0 est la valeur de $\rho(t)$ à l'instant initial $t = 0$.

Il nous faut donc calculer la valeur de $\exp(Qt)$. Une manière simple de calculer cette quantité est de diagonaliser Q, c'est à dire de trouver deux matrices A et D satisfaisant la relation $Q = ADA^{-1}$ où D est une matrice diagonale. Dans ce cas, on peut montrer que:

$$e^{Qt} = Ae^{Dt}A^{-1}. \tag{1.14}$$

Comme D est une matrice diagonale, on a $(e^D)_{aa} = e^{(D)_{aa}}$, et on peut donc facilement calculer e^D. Ainsi, on dispose d'une formule simple permettant de calculer la probabilité qu'un nucléotide soit présent à un instant t en fonction de son état à l'instant initial, et des taux de mutation. Il nous reste donc à définir un modèle pour ces taux de mutation.

L'hypothèse la plus simple consiste à considérer que toutes les mutations se produisent avec la même probabilité $q/4$. Dans ce modèle, connu sous le nom de modèle de Jukes et Cantor [15], la matrice Q est définie par:

$$Q = q \begin{pmatrix} -3/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & -3/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & -3/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & -3/4 \end{pmatrix}. \tag{1.15}$$

Après diagonalisation de la matrice Q, on peut appliquer l'équation (1.14) et on trouve alors pour $P(t)$:

$$P_{ab}(t) = \begin{cases} \dfrac{1}{4}(1 - e^{-qt}), & \text{for } a \neq b \\ \dfrac{1}{4} + \dfrac{3}{4}e^{-qt}, & \text{for } a = b. \end{cases} \tag{1.16}$$

En utilisant les mêmes arguments, on peut aussi résoudre des modèles plus généraux dans lesquels les taux de mutations varient d'une substitution à l'autre.

1.1.1.3 La seconde règle de parité de Chargaff

Une quinzaine d'année après la découverte de la structure de l'ADN, Chargaff et ses collègues ont observé une autre propriété intéressante des séquences d'ADN. Après avoir séparé les deux brins d'ADN du génome d'une bactérie, *Bacillus subtilis*, ils ont observé que sur chacun des deux brins, les proportions en A et T d'une part, et les proportions en G et C de l'autre étaient, là aussi, à peu près égales [16], de sorte que:

$$\begin{cases} n_A \sim n_T \\ n_C \sim n_G \end{cases} \tag{1.17}$$

où n_a représente ici le nombre de bases a sur un brin d'ADN.

Cette propriété (que l'on appelle la seconde règle de Chargaff, ou PR2) a été observée chez la majorité des génomes connus [17, 18]. À la différence de la première loi de Chargaff toutefois, des exceptions à cette règle ont été identifiées, notamment dans des génomes de mitochondrie [18]. À la fin des années 1990 , Lobry et Lobry [19] ont démontré mathématiquement que lorsque les mutations affectent les deux brins de la même manière (condition appelée "no strand bias condition" [20]), les séquences évoluent vers un état d'équilibre dans lequel la seconde loi de Chargaff est respectée. Comme nous l'avons vu précédemment, lorsqu'une mutation se produit et n'est pas réparée, elle affecte les deux brins de la séquence. Par exemple, si un A est remplacé par un C sur l'un des deux brins, alors un T va être remplacé par un G à la même position sur le brin complémentaire. Ainsi, si les mutations affectent indifféremment sur les 2 brins, les taux associés à une mutation de A vers C et de T vers G seront égaux. De ce fait, chaque mutation ayant une mutation équivalente sur le brin complémentaire, six taux de mutation sont suffisants pour décrire l'évolution d'une séquence. Dans ce cas, la

matrice des taux instantanés est donnée par:

$$Q = \begin{pmatrix} \bullet & \mu_{GT} & \mu_{CT} & \mu_{AT} \\ \mu_{AC} & \bullet & \mu_{CG} & \mu_{AG} \\ \mu_{AG} & \mu_{CG} & \bullet & \mu_{AC} \\ \mu_{AT} & \mu_{CT} & \mu_{GT} & \bullet \end{pmatrix} \tag{1.18}$$

où $\bullet$ est encore une fois défini de manière à ce que la somme de toutes les valeurs d'une colonne soit nulle. On peut montrer que l'évolution d'une séquence selon une chaîne de Markov utilisant une matrice de taux instantanés prenant la forme précédente évolue vers un état d'équilibre dans lequel la seconde règle de parité de Chargaff est toujours vérifiée [19]. Comme cette règle est vérifiée dans la plupart des génomes, on peut en conclure que, dans la majorité des espèces, les taux de mutation sont les mêmes sur les deux brins de l'ADN.

Toutefois, cette seconde règle n'est pas aussi absolue que la première, notamment parce qu'elle correspond à un état d'équilibre de la séquence. Ainsi, il a été observé que dans certains génomes, et plus particulièrement dans des petites régions de certains génomes, cette règle n'était pas validée, indiquant que dans ces régions, les mutations n'affectent pas les deux brins de manière équivalente [21–24]. Par la suite, l'étude des dérogations à PR2 ont permis de mettre en évidences des caractéristiques variées de certaines régions des génomes [25], comme par exemple la position des origines de réplications [26].

La seconde loi de Chargaff constitue ainsi un exemple intéressant qui démontre l'intérêt de la modélisation en biologie. Dans un premier temps, expliquer des propriétés statistiques simples permet d'appréhender des mécanismes biologiques généraux se déroulant à l'échelle globale. Dans un second temps, les modèles simples peuvent servir de cadre général à partir desquels on peut identifier des déviations, notamment à des échelles plus localisées. L'analyse de ces déviations peut alors permettre d'identifier des phénomènes nouveaux et de de développer une vision plus précise des processus biologiques qui façonnent les objets biologiques.

1.1.1.4 Distribution des Longueurs d'Appariement

Dans cette thèse, nous allons étudier différents mécanismes biologiques qui engendrent de longues répétitions dans les génomes, mécanismes qui ne sont pas pris en compte dans les modèles que nous avons présentés jusqu'à maintenant. Afin d'étudier ces mécanismes, nous allons analyser la distribution de la longueur des appariements exacts (c'est à dire des segments d'ADN partageant exactement la même séquence), qui sont maximaux (qui ne sont pas inclus dans des appariements plus grands) que l'on notera $M(.)$. On peut calculer ce type de distribution soit à partir de la comparaison d'un génome avec lui-même, soit à partir de la comparaison des génomes de deux espèces distinctes.

Si l'on calcule cette distribution pour des séquences aléatoires obtenues à l'aide du modèle iid, on peut établir que cette distribution est géométrique, et est donnée par:

$$M_{\text{iid}}(r) = \frac{1}{2} L_1 L_2 (1-p)^2 p^r, \tag{1.19}$$

où L_1 et L_2 sont les tailles des deux génomes à comparer et où p représente la probabilité que deux nucléotides choisis aléatoirement soient identiques, d'où:

$$p = f_A^{(1)} f_A^{(2)} + f_C^{(1)} f_C^{(2)} + f_T^{(1)} f_T^{(2)} + f_G^{(1)} f_G^{(2)}, \tag{1.20}$$

avec $f_a^{(i)}$ la fréquence du nucléotide a dans la séquence i. On pourra par ailleurs remarquer que le maximum de cette fonction, (qui correspond au plus long appariement attendu entre deux séquences aléatoires), suit la loi de Gumbel qui est utilisée pour tester la probabilité qu'un appariement entre deux séquences soit dû au hasard [27, 28].

Toutefois, la distribution que l'on obtient lorsque l'on réalise la même expérience sur des génomes eucaryotes (que l'on compare des génomes entre eux ou que l'on compare un génome avec lui-même) est nettement différente de la distribution théorique. Plus précisément, pour de vrais génomes, on obtient largement plus d'appariements de grande taille que ce à quoi l'on s'attend selon le modèle iid. Plus surprenant, on observe que $M(r)$, le nombre d'appariements obtenus en fonction

de leur taille r, est distribué selon une loi de puissance, c'est à dire que $M(r) = \mathscr{C}r^{\alpha}$ où α est l'exposant de la distribution, et $\mathscr{C}$ une constante de normalisation. L'exposant α de cette distribution varie en fonction des cas, et le but de cette thèse sera d'expliquer mathématiquement ces observations et de montrer comment différents mécanismes biologiques, et notamment de duplications, sont à l'origine de ces différentes distributions.

1.1.2 Des processus d'évolution plus complexes

Jusqu'à présent, nous avons discuté des modèles dont les principes généraux étaient relativement simples, ne prenant en compte que les propriétés les plus basiques du fonctionnement de l'ADN, et négligeant de nombreux processus plus complexes, dont l'importance au cours de l'évolution des espèces est aujourd'hui pleinement reconnue (bien que pas toujours totalement élucidée). Dans cette partie, nous allons présenter certains de ces processus, que nous étudierons dans la suite de cette thèse.

1.1.2.1 Les éléments transposables

Les éléments transposables (ET), que l'on désigne aussi sous le nom d'éléments répétés, ou de transposons, sont de petites séquences d'ADN, dont la taille va de quelques centaines de paires de bases à plusieurs kilopaires de bases (kbps), et qui ont la capacité de se dupliquer par elle-même dans les génomes. Ces éléments constituent une part importante (et fortement variable d'une espèce à l'autre) des génomes eucaryotes. On estime par exemple qu'environ 50% du génome humain est constitué de ce type d'éléments [29], tandis qu'ils représentent environ 85% du génome du maïs [30] et seulement 14% du genome de l'organisme modèle *Arabidopsis thaliana* [31].

Il a été établi que les duplications de transposons se déroulent par vagues, chacune s'étendant sur un court laps de temps pendant lequel un très grand nombre de duplications se produisent, jusqu'au moment où l'organisme hôte développe un

mécanisme empêchant d'autres duplications de se produire. À partir de ce moment là, les séquences de ces transposons restent inertes dans le génome et évoluent, accumulant des mutations progressivement, comme tout le reste de l'ADN non codant. Il existe de nombreux types de transposons différents, qui sont classifiés en familles partageant le même mode de duplication, ou des analogies de séquence. La famille d'éléments répétés la plus représentée (en terme de nombre de copies) dans le génome humain est la famille des éléments Alu [32, 33], un petit transposon d'environ 300 paires de base dont on peut trouver plus d'un million de copies dans le génome humain (pour un aperçu complet des connaissances concernant ces éléments, voir Batzer and Deininger [34] ou Deininger and Batzer [35]).

La plupart du temps, les ETs sont considérés comme des séquences égoïstes, envahissant le génome de leur hôte, et dont les insertions ont des conséquences délétères. Notamment si une de ces insertions se produit au sein d'un gène, la séquence de ce gène va être perturbée, risquant ainsi de le rendre non fonctionnel. Par ailleurs, on les considère aussi souvent comme des vecteurs de l'instabilité des génomes [36].

Toutefois, de nombreuses études ont également souligné les effets positifs de l'insertion de ces éléments pour les génomes hôtes. Par exemple, il a été mis en évidence que près de 25% des promoteurs (des petites régions situées au début d'un gène permettant à l'ARN polymérase de se fixer à l'ADN, et ainsi au gène de s'exprimer) du génome humain contiennent des séquences apparentées à des ETs [37, 38]. Dans le même ordre d'idées, 7.8% des gènes exprimés dans le génome d'*Arabidopsis thaliana* contiennent une région dont la séquence partage un haut degré de similarité avec la séquence d'un transposon [39]. Par ailleurs, certaines vagues de duplications d'ETs ont été associées à des évènements de spéciations (une insertion massive de ces éléments a par exemple eu lieu au moment de la formation des primates [40]), ainsi qu'à des mécanismes d'évolutions adaptatifs. Il a par exemple été observé dans le génome de certaines plantes (et dans une moindre mesure chez certaines drosophiles) une nette augmentation du taux d'insertions d'ETs après qu'elles aient été soumises à des conditions de stress important [37]. Toutefois, des

preuves directes reliant une vague de duplications d'ETs à une vague d'innovations dans les génomes n'ont pas été identifiées.

Enfin, la séquence, ainsi que le mode de duplication des transposons varient fortement d'une espèce à l'autre, et il est donc très probable qu'un grand nombre d'éléments de ce type (particulièrement dans les génomes d'espèces n'ayant pas fait l'objet de recherches approfondies) soient encore à découvrir. Le séquençage des génomes de nouvelles espèces, ainsi que l'amélioration de la qualité des génomes déjà séquencés à l'aide de nouvelles technologies va ainsi permettre dans les années à venir de mieux comprendre l'histoire évolutive de ces éléments, ainsi que leur impact sur l'évolution des génomes.

1.1.2.2 Les répétions possédant un petit nombre de copies

Les duplications de gène — À la différence des transposons qui sont de petites séquences répétées un nombre très élevé de fois, les génomes possèdent également des séquences dupliquées un petit nombre de fois. Ces séquences, qui ne possèdent pas, elles, la capacité de se dupliquer par elles-mêmes, sont le résultat d'erreurs ayant lieu à différents stades de la vie d'une cellule. On appelle deux séquences homologues issues d'un tel événement des paralogues (par opposition aux orthologues, qui sont des séquences homologues résultant de la divergence de deux espèces).

L'importance évolutive de ces duplications à petit nombre de copies (DPNCs), qui a été popularisé par l'intermédiaire du livre de Susumu Ohno *Evolution by Gene Duplication* [41] est aujourd'hui largement reconnue. Dans ce livre, Ohno développe l'idée qu'une seule copie d'un gène est suffisante pour remplir une fonction. Ainsi, lorsqu'un gène est dupliqué, les contraintes empêchant la séquences d'accumuler des mutations ne restreignent l'évolution que d'une seule des deux copies seulement, de sorte que la seconde peut accumuler des mutations librement. La plupart du temps, les mutations affectant le plus libre des deux gènes auront pour conséquence la perte de fonction du gène en question (on parle alors de pseudogénisation). Toutefois, il peut arriver que ces mutations permettent aux gènes

d'acquérir une nouvelle fonction. Le destin des gènes à la suite d'un événement de duplication est une question très complexe, et qui fait encore aujourd'hui l'objet de nombreuses études. Leur devenir peut être décrit par l'un des trois scénario suivants [42, 43]:

– Le premier scénario possible est la pseudogénisation d'une des deux copies, qui va accumuler des mutations jusqu'à devenir totalement inactivée, tandis que l'autre copie restera inchangée.

– Dans un second scénario, appelé néofonctionnalisation, l'une des deux copies acquiert une nouvelle fonction, tandis que l'autre continue à remplir la fonction originale du gène avant la duplication.

– Le dernier scénario possible est la subfonctionnalisation des deux gènes. Dans ce cas, les deux copies perdent une partie de leur fonction à la suite de mutations, jusqu'à atteindre le point où les fonctions conjuguées des deux gènes seront équivalentes à celle du gène unique avant la duplication. Un cas particulier de subfonctionnalisation se produit lorsque le gène ancestral possède plusieurs fonctions. Dans ce cas, chacune des copies peut se spécialiser et améliorer l'une de ces fonctions (et notamment aux dépens d'une autre fonction qui est maintenue par la seconde copie).

Dans la suite, nous allons présenter trois mécanismes biologiques différents ayant pour conséquence la formation de DPNCs et que l'on étudiera au cours de cette thèse.

Les duplications segmentaires — Les duplications segmentaires (DSs) sont classiquement définies comme de longs ($> 1kb$) segments d'ADN partageant une identité de séquence très élevée ($> 95\%$) [44]. Ces segments constituent une part importante des génomes eucaryotes (5.5% du génome humain et 5% du génome de la souris par exemple [44, 45]). Ces duplications peuvent avoir lieu à la suite de nombreux mécanismes biologiques différents, et sont, le plus souvent, la conséquence d'une erreur ayant lieu suite à des cassures double brin (ou "double-strand breaks") de l'ADN. Un de ces mécanismes nécessite la recombinaison de

segments non homologues de l'ADN, et se produit ainsi par l'intermédiaire de duplications pré-existantes (par exemples d'autres DSs, ou bien même des éléments répétés tel que ceux de la famille Alu [46, 47]), et la probabilité qu'un événement de ce type se produise est proportionnelle au degré de similarité existant entre les deux copies servant d'intermédiaire. D'autres mécanismes générant des duplications ont été mis en évidence, telles que les jonctions non homologues [48, 49] (qui se produisent principalement dans les télomères) ou encore les FoSTES (pour "fork stalling and template switching"). Pour un aperçu plus complet de ces différents mécanismes, voir Hastings, Lupski, Rosenberg, and Ira [50].

Comme les DSs sont, la plupart du temps, des effets secondaires ayant lieu à la suite d'erreurs dans les mécanismes de réparation de l'ADN, ces duplications ne sont pas distribuées équitablement le long des génomes, et sont sur-représentées dans les régions instables. Ainsi, les centromères, et dans une moindre mesure les télomères sont particulièrement enrichis en DSs [49]. Par exemple, 31% de tous les nucléotides dupliqués du génome humain sont situés à moins de 5 Mbps des centromères, ce qui est 6 à 7 fois plus que ce qui serait attendu dans le cas où les duplications seraient réparties équitablement dans le génome [51].

Enfin, il est intéressant de noter que, même si les duplications de gènes font l'objet d'un grand nombre d'études, les duplications ne ciblent pas prioritairement les gènes. Notamment, la fréquence avec laquelle une séquence est dupliquée n'est que faiblement corrélée au fait que cette séquence contienne un gène. Cette corrélation est par ailleurs plus probablement une conséquence d'un plus fort taux de rétention des DSs contenant des gènes plutôt que d'un taux plus élevé de duplication dans les régions contenant un grand nombre de gènes.

Rétroduplications — Un autre processus bien connu induisant des répétitions dans les génomes est le processus de rétroduplication. Dans ce cas, une molécule d'ARNm est tout d'abord retrotranscrite en ADN, et est ensuite intégrée dans le génome. Les duplications qui en résultent ont des propriétés différentes des DSs.

Tout d'abord, la plupart des gènes sont constitués d'une alternance de séquences codantes (les exons) et de séquences non-codantes (introns). Ainsi, bien que le gène tout entier soit transcrit en ARN, après l'épissage, les introns sont exclus, et l'ARNm mature n'est constitué que d'exons. Cet ARNm mature est ainsi nettement plus court que le gène entier, de sorte que les retroduplications produisent des duplications partielles qui sont, en moyenne, plus courtes que les DSs.

Par ailleurs, les duplications ainsi produites ne ciblent que les exons, et, de ce fait, copient pas les séquences régulatrices (et notamment les promoteurs), qui sont indispensables au fonctionnement d'un gène (et à sa transcription en particulier). Ainsi, les séquences produites par rétroduplication sont, la plupart du temps, des séquences non codantes hautement similaires à des gènes et sont appelées des pseudogènes processés [52, 53]. Toutefois, il a été démontré que certains de ces pseudogènes étaient effectivement fonctionnels de sortes que leur rôle dans les génomes est toujours débattu, voir notamment Kaessmann, Vinckenbosch, and Long [53] et Okamura and Nakai [54]. Mais il semble bien que, dans la majorité des cas, les pressions de sélection s'appliquant à ces séquences ne sont pas suffisantes à leur maintien dans les génomes, et leur destin consiste à disparaitre par l'accumulation de mutations délétères.

Enfin, comme nous l'avons vu, seuls les exons peuvent-être rétrodupliqués (tandis que les DSs peuvent cibler n'importe quelle région du génome). Il a même été démontré que la probabilité qu'une molécule d'ARNm soit rétrodupliquée est proportionnelle à son niveau d'expression (et donc à la quantité de cette molécule d'ARNm dans la cellule) [55, 56].

En conséquence, à la différence des SDs, où les deux copies sont globalement équivalentes, lors de rétroduplications, il est possible de distinguer la séquence originale (ou séquence mère) de celle qui a été dupliquée (séquence fille), et ces deux séquences ont des propriétés nettement différentes. En particulier, après la duplication, la séquence mère pourra être l'objet d'autres duplications, mais cela ne sera pas le cas de la séquence fille (parce qu'elle n'est pas transcrite).

Les duplications de génome entier — Les deux processus que nous venons de décrire affectent les génomes relativement fréquemment, et ciblent un petit nombre de nucléotides. Il arrive toutefois que des duplications à de nettement plus grandes échelles se produisent, entrainant par exemple la duplication d'un chromosome entier, voir de tout le génome d'un individu (DGE).

Ces événements, ont, bien évidemment, des conséquences beaucoup plus importantes pour les génomes que ceux décrits précédemment, et sont pour cette raison beaucoup plus rares. Mais lorsqu'ils se produisent, ils ont pour conséquence la mise à disposition d'une vaste quantité d'ADN plus ou moins libre d'évoluer, permettant ainsi des innovations génétiques, et ils pourraient être à l'origine de nombreux événements de spéciations [57]. Par exemple, 15% (resp. 31%) des événements de spéciations des plantes à fleur (resp. des fougères) seraient associés à des DGE [58]. Il a par ailleurs été observé que dans la plupart des cas, à la suite de ces événements, on observait de nombreuses délétions dans les génomes [59], ce qui a pour conséquence de compliquer l'étude et la mise en évidence de ces événements. Notamment, il a été postulé il y a plus de 40 ans que deux événements de DGE auraient eu lieu à l'origine de la lignée des vertébrés, mais le débat autour de cette question dans la communauté scientifique reste toujours ouvert.

1.1.3 Plan de la Thèse

Le but de cette thèse est d'étudier la distribution des tailles des répétitions dans les génomes eucaryotes. Le chapitre 2, détaille certaines de nos expériences, ainsi que des modèles mathématiques que nous utiliserons par la suite. Dans le chapitre 3, nous nous intéresserons aux cas des comparaisons de génomes avec eux-même. Nous commencerons par expliquer le cadre mathématique que nous utiliserons ensuite tout au long de cette thèse, l'appliquerons à différents mécanismes de duplication, et discuterons les implications des déviations par rapport à notre modèle observées dans les génomes de deux primates. Le chapitre 4 porte sur l'analyse des distributions obtenues en comparant différents génomes, et aura pour but de montrer que les distributions obtenues dans ce cas sont la conséquence des

variations du taux de mutation le long des génomes. Au cours du chapitre 5, nous nous intéresserons aux conséquences des duplications de génomes entiers, qui nécessitent de faire appel aux résultats des deux chapitres précédents. Enfin, dans le chapitre 6, nous étudierons le cas particulier des comparaisons des séquences codantes des génomes, et montrerons que les modèles développés jusqu'ici ne permettent pas d'expliquer les résultats obtenus dans ces cas là.

La majeure partie des résultats présentés au cours de cette thèse ont été publiés dans des revues scientifiques à comité de lecture. Le modèle le plus simple, qui ne s'intéresse qu'aux duplications segmentaires (Section 3.2) est l'objet d'un article publié dans *Physical Review Letters* [60]. La généralisation à d'autres mécanismes de duplication, (Section 3.3 et 3.4) ainsi qu'aux comparaisons de génomes (Section 4.4.1) est présenté dans un article du journal *Molecular Biology and Evolution* [61]. Enfin, les calculs relatifs aux propriétés des arbres de Yule est l'objet d'un article dans *Plos One* [62]. Les détails développés dans cette dernière publication sortent du cadre de cette thèse. Pour cette raison, seuls quelques résultats de cette publication sont discutés dans le présent document, et l'article est présenté en appendice (voir Appendice C).

1.2 English Version of the Introduction

In this thesis, we study the length distribution of maximal repeats in eukaryotic genomes, and more generally the length distribution of maximal exact matches between genomes of different species. Indeed, these distributions strongly deviate from what one could expect from simple probabilistic models and present a power-law behavior. We will show that simple evolutionary models are able to account for these deviations. In the Introduction, we first review some simple statistical properties of DNA sequences, and show that deciphering these properties have initiated significant progress in the field of genetics and evolution. The scientific approach we developed is inherited from these seminal historical studies. We will then introduce some biological processes and mathematical concepts that will be studied more specifically later on.

1.2.1 Statistical Properties of Genomes

The genome of an organism is defined as the entire genetic material it carries. It contains all the information that an individual needs to develop from a single cell, to grow and to reproduce. This information is transmitted from an organism to its progeny during reproduction. The genetic information is encoded in a long molecule, the Deoxyribonucleic Acid (DNA), which is a polymer of four different monomers called nucleotides. Each of these monomers are composed of a sugar, a phosphate group and a nitrogenous base. The sugar and the phosphate groups are the same in all four nucleotides, but there are four possible nitrogenous bases. These four bases (Adenine (A), Guanine (G), Cytosine (C), Thymine (T)) are classified into two groups of molecules, the purines (A and G), and the pyrimidines (C and T). Bases belonging to the same group share a high chemical similarity. The purine bases are composed of two aromatic compounds while the pyrimidine bases are composed of only one aromatic cycle, and are thus smaller (see Fig. 1.2 for the full chemical structure of the four nucleotides).

Cells of all living organisms possess the ability to interpret the complex information encoded in the DNA to produce RNA molecules (via a process called transcription) that are then translated into proteins, which in turn perform the essential functions of living cells. For this reason, DNA is often described as the cookbook of an organism, written in an alphabet of four letters.

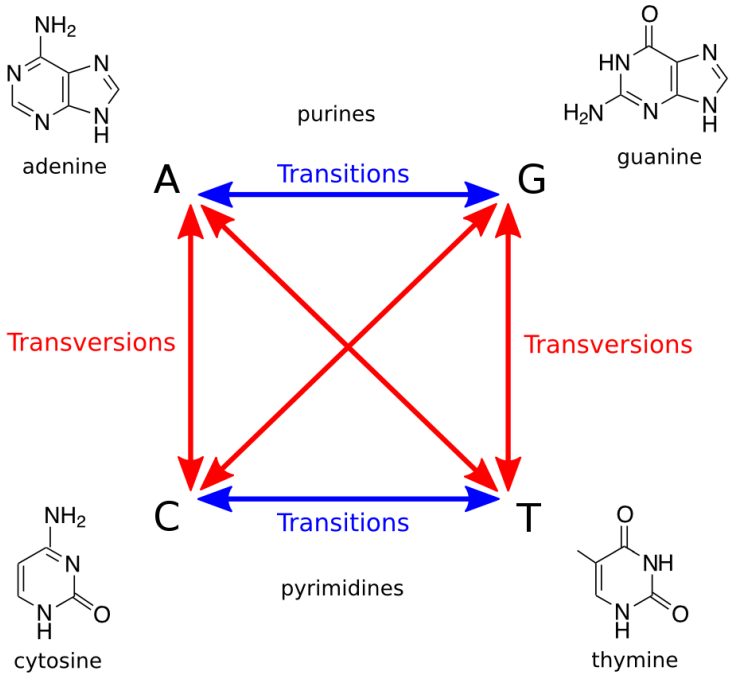

FIGURE 1.2: The classification of mutations into transversions (in red) and transitions (in blue). The full chemical structure of the four nucleotide bases are also represented. Picture taken from the Rosalind website (http://www.rosalind.info).

1.2.1.1 Chargaff's First Parity Rule

Early on in the history of genetics, while it had not yet been established that genetic information is encoded in the DNA, scientists started to study statistical features of DNA, in order to understand its properties.

A basic but yet historically very important property has been highlighted by Erwin Chargaff. Chargaff studied the proportion of each nucleotide in the DNA of the cell of many species, and found what is today known as Chargaff's first parity rule

[1]:

$$\begin{cases} \dfrac{n_A}{n_T} = 1 \\ \dfrac{n_C}{n_G} = 1 \end{cases} \tag{1.21}$$

where n_a stands for the number of nucleotides a. Examples in different species from Chargaff's experiment are shown in Table 1.2. This property — together with X-ray pictures of DNA that we owe to Franklin and Gosling [2] — led to the discovery of DNA structure by Watson and Crick [3]. Watson and Crick's model for the structure of DNA states that DNA molecules (or strands) are linked by pairs, and that each nucleotide on one strand is associated to a nucleotide on the complementary strand. This link is made possible by chemical interactions between nucleotides. Due to steric constraints, purine-purine and pyrimidine-pyrimidine association are much less stabilizing than purine-pyrimidine pairing. The four possible associations left are thus A - T, A - C, C - G, and C - T. But stabilizing chemical interactions (π-stacking and hydrogen bounds [4]) preferentially occur between A and T on the one hand and C and G on the other. Thus, in double stranded DNA, A are always paired to T and C are always paired to G. These preferential associations result in the fact that in each cell, the proportions of A and T as well as proportions of C and G are always equal, and thus to Chargaff's first parity rule.

Many fundamental processes involving DNA rely on the association of DNA strands by pairs. Thus, its discovery pave the way to a better understanding of how life works. Due to the pairing rule, both associated strands are complementary and contain the same information, such that the knowledge of one of the two strands is enough to fully reconstruct its complementary strand. Thanks to this property, DNA can also easily replicate. The cell machinery first separates the two complementary strands and each of them then serves as a template to produce a new DNA molecule. At the end of the process, the cell contains two identical pairs of DNA, and it can thus divide into two daughter cells containing the same DNA content.

DNA replicates with a very high fidelity. Sometimes however, errors occur during

	% of bases				Ratios		
Source	A	G	C	T	A/T	G/C	%GC
ψX174	24.0	23.3	21.5	31.2	0.77	1.08	44.8
Maize	26.8	22.8	23.2	27.2	0.99	0.98	46.1
Octopus	33.2	17.6	17.6	31.6	1.05	1.00	35.2
Chicken	28.0	22.0	21.6	28.4	0.99	1.02	43.7
Rat	28.6	21.4	20.5	28.4	1.01	1.00	42.9
Human	29.3	20.7	20.0	30.0	0.98	1.04	40.7

TABLE 1.2: This table represents a sample of Chargaff's 1952 data, listing the nucleotide composition of DNA of several organisms. Table reproduced from Bansal [5].

the replication process. For instance, one base can be mistakenly inserted in place of another. As a result, the newly produced double stranded molecules are exactly identical to the original molecule at all positions (or loci) but one. This event is called a point mutation. As point mutations can result of the change of any of the four possible nucleotide in any of the three others, there are twelve possible mutations. Although rare and sources of deleterious effects, these errors also allow species to evolve and to adapt to their changing environment. These changes in DNA do not only happen during replication, and the alteration of a nucleotide can also occur at another step of the cell life. And since nucleotides belonging to the same chemical group are more closely related to one-another than to the two others, the four mutations that change a purine into another purine or a pyrimidine into another pyrimidine (called transitions, see Fig. 1.2) occur much more often than the eight others (called transversions).

When a mutation occurs on one strand, a newly inserted nucleotide and its homologous base on the other strand are not anymore paired according the pairing rule of Watson and Crick's model. This results in an alteration of the structure of the DNA, that can be identified by the repair machinery present in the cell, whose function is to repair the error. However, most of the time, the cell has no way to identify which base is in the unaltered state, and which base has been changed.

Thus, one of the two bases is randomly replaced so that the two bases are properly associated. Then, half of the time, the mutation is repaired, and half of the time the mutation gets fully integrated.

1.2.1.2 Modelling DNA Sequences

Different individuals belonging to the same species share very similar genomes. For this reason, one can build a reference genome for each species, representing the typical genome of one individual. The size of genomes is highly variable from one species to another, ranging from several hundreds of kilobase pairs (kbps), to hundreds of gigabase pairs (Gbps), while the size of typical mammalian genomes is of the order of several Gbps (3.2 Gbps for the Human). To give a comparison, the french Wikipedia was composed of roughly 4.2 billion letters in total in June 2015 according to wikipedia itself. Unlike Wikipedia however, the entire Human genome (as well as other eukaryotic genomes) cannot be directly interpreted. Indeed, only a small percentage of eukaryotic genomes codes for proteins (this proportion is of the order of 1% in the Human genome), while most of the genome (the non-coding DNA) is thought to be mostly non-functional (although the potential function of the non-coding DNA is the subject of a lively and passionated debate in the scientific community [6–8]). This explains why the size of genomes varies so much from one species to another, and why the complexity of an organism does not correlate well with the size of its genome.

Studying objects of such a large size and composed of repeated units (here nucleotides) allows to perform statistical analysis. One of the first statistical analysis performed on genomes aimed at identifying short sequences of nucleotides (or words) that were either exceptionally frequent or rare, with the idea that these exceptional events would be the signature of a biological significance of these words [9–11]. Words of particular biological importance, such as for instance motifs recognized as transcription start sites by transcription factors are expected to be overrepresented in the sequence, while certain types of motif might be deleterious, and thus avoided. But in order to identify exceptional words one first needs

to develop a random model, to differentiate events that occur "by chance" from biologically meaningful events.

Random sequences — One can model a DNA sequence $\mathcal{S} = (s_1, ..., s_L)$ as a chain of L letters, where each letter belongs to the four letter alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. The simplest possible way to model a DNA sequence, is the random independent and identically distributed (iid) model. In this framework, all base pairs s_i are independent from each other, and the probability of each nucleotide to be observed is constant at all positions of the sequence. The parameters of the model are the length of the sequence L, and the frequency of each of the four base pairs f_A, f_T, f_C and f_G. In that case we have:

$$P(s_i = a) = f_a, \text{ with } a \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}, \ \forall i \in 1, \ldots, L. \tag{1.22}$$

k-letter words W are defined as subsequences of k consecutive letters (and are for this reason also termed k-mers). In a sequence of size L, there are $L-k+1$ words of length k, and the word starting at position i is defined as $W_i = (s_i, \ldots, s_{i+k-1})$. The probability of a given word $W = (w_1, ..., w_k)$ of length k to appear in a sequence of size L at the position i is then:

$$P(W_i = W) = \prod_{j=1}^{k} P(s_{i+j-1} = w_j) = \prod_{j=1}^{k} f_{w_j}, \tag{1.23}$$

from which it follows that a word of length k appears in a sequence of length L on average n_k times, with

$$n_k = (L - k + 1) \prod_{j=1}^{k} f_{w_j}. \tag{1.24}$$

Markov Models — *This section borrows heavily from Robin, Rodolphe, and Schbath [12], where one can find broader developments of the topic discussed below.*

More sophisticated models have been developed and have been shown to be powerful tools to understand biological processes that shape genomes. One such class

of models which has been widely used are Markov chain models. Such models have been applied to a broad range of fields (for general discussions about Markov Chains, see Karlin and Taylor [13] for instance). The most simple Markov chains (also called first order Markov chains) are processes where events at step i only depend of the state of the chain at step $i-1$, and is independent from all other previous states.

In the case of DNA, the value of each letter s_i depends only on the value of its left neighbor s_{i-1}. There are 16 possible couples of letters (s_{i-1}, s_i), and thus 16 transition probabilities. The transition probabilities $p_{a\to b}$ from a letter $a \in \mathcal{A}$ to letter $b \in \mathcal{A}$ is defined as

$$p_{a\to b} = P(s_i = b | s_{i-1} = a). \tag{1.25}$$

These probabilities can be estimated from an observed DNA sequence using the maximum likelihood method. The likelihood of a model is defined as the probability to obtain a set of data given the model. Then, the values of the parameters that maximize the likelihood give accurate estimators of the parameters.

In the present case, the maximum likelihood estimators (MLE) of the transition probabilities are:

$$p_{a\to b} = \frac{n_{ab}}{\sum_{c\in\mathcal{A}} n_{ac}}, \tag{1.26}$$

where n_{ab} is defined as the count of the 2-letter word $W = ab$. In that sense, this model is an extension of the iid model where the parameters were the single nucleotide frequencies only. Following a similar procedure, one can build $m-$order Markov chains, where the letter s_i depends on the m previous letters. In this case, one has to define 4^m transition probabilities, whose MLE will depend on the frequencies of words of length $m+1$. Thus, the number of parameters increases with the order of the Markov chain. Hence, the higher the order of the Markov chain, the more accurately it describes the DNA sequence.

However, regarding modeling, more accurate does not always mean better. Recall that one of the motivation to model DNA was to produce a random model in

order to differentiate exceptional words from those occurring just by chance. For instance, one cannot expect to find exceptional words of length k in a sequence using a k-order Markov chain model. An extreme case of evidently pointless model consists in representing a sequence of length L using a $L-1$-order Markov chain.

Heterogenous Markov Chains — So far, we only described homogeneous Markov chain models. In these models, transition probabilities are the same all along the sequence. But we have already seen that only a small fraction of eukaryotic genomes is coding for proteins. As the coding potential of a region constraints its statistical properties, such features are not taken into account in homogeneous Markov chains. To deal with these irregularities, heterogeneous Markov chains have been introduced. In these models, the transition probabilities are different from one region to another. One subclass of heterogeneous Markov chains of particular interest for DNA modeling are the hidden Markov models. These models have been widely used, notably to detect coding regions in genomes [14].

Modeling the Evolution of Sequences — One can also use Markov chains to model the evolution of DNA in time. In these models, it is assumed that each site evolves independently. In this case, it is realistic to use a first order Markov model, that is, to consider that the mutation rate of a nucleotide depends on the present letter only, and not on the letter that could be found at the same position in the past. Indeed, in real genomes, informations regarding the nucleotides that could be found at given positions in the history of a species are not stored.

To model the evolution of a sequence $\mathcal{S}(t) = (s_1(t), ..., s_L(t))$ in time, one uses a continuous time Markov chain. In such a model, the probability $\rho_a(t)$ that a nucleotide $s_i(t) = a$ is given by a Master equation:

$$\frac{\delta}{\delta t}\rho_a(t) = \sum_{b \neq a} \rho_b(t) Q_{ba} - \rho_a(t) \sum_{b \neq a} Q_{ab} \tag{1.27}$$

where $(a, b) \in \mathcal{A}^2$ are nucleotides and Q_{ab} is the substitution rate from state a to state b such that the probability that a nucleotide a mutates to b in an infinitesimal

small time δt is $Q_{ab}\delta t$. One can write these equations using matrices and find:

$$\frac{\delta}{\delta t}\rho = \rho(t)Q \tag{1.28}$$

where $\rho(t)$ is a row vector of dimension 4 :

$$\rho(t) = \Big(\, \rho_A(t),\, \rho_C(t),\, \rho_G(t),\, \rho_T(t)\Big) \tag{1.29}$$

and Q a 4×4 matrix defined as:

$$Q = \begin{pmatrix} \bullet & Q_{CA} & Q_{GA} & Q_{TA} \\ Q_{AC} & \bullet & Q_{GC} & Q_{TC} \\ Q_{AG} & Q_{CG} & \bullet & Q_{TG} \\ Q_{AT} & Q_{CT} & Q_{GT} & \bullet \end{pmatrix} \tag{1.30}$$

where the diagonal terms denoted by $\bullet$ are defined such that columns sum up to zero i.e.:

$$Q_{aa} = -\sum_{a \neq b} Q_{ab}. \tag{1.31}$$

The solution of these differential equations is given by:

$$\rho(t) = \rho_0 P(t) \tag{1.32}$$

with

$$P(t) = \exp(Qt) = \sum_{n=0}^{\infty} \frac{(Qt)^n}{n!}, \tag{1.33}$$

where $P_{ab}(t)$, the ab^{th} coordinate of $P(t)$ is the probability that a site changes from state a to state b during a finite time interval t, and where ρ_0 is the value of ρ at the initial time $t = 0$.

An easy way to calculate the value of $\exp(Qt)$, is to diagonalize Q, that is, to find two matrices A and D such that $Q = ADA^{-1}$, D being a diagonal matrix. Then, one obtains:

$$e^{Qt} = Ae^{Dt}A^{-1}. \tag{1.34}$$

As D is diagonal, it follows that $(e^D)_{aa} = e^{(D)_{aa}}$, which is easy to calculate.

Using this framework, the simplest possible model is the Jukes and Cantor model [15], where all mutations occur with the same probability $q/4$, such that the matrix Q is defined as:

$$Q = q \begin{pmatrix} -3/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & -3/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & -3/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & -3/4 \end{pmatrix}. \tag{1.35}$$

Diagonalizing Q using Eq. (1.34), we find for the probability matrix $P(t)$:

$$P_{ab}(t) = \begin{cases} \frac{1}{4}(1 - e^{-qt}), & \text{for } a \neq b \\ \frac{1}{4} + \frac{3}{4}e^{-qt}, & \text{for } a = b. \end{cases} \tag{1.36}$$

1.2.1.3 Chargaff's Second Parity Rule

Later on, Chargaff and his coworkers separated the two strands of the DNA of the model bacterium *Bacillus subtilis*, and calculated the proportions of each base pair independently in the two strands. They found that even in a single strand, the proportion of A and T on the one hand, and the proportion of G and C on the other were approximately equal [16] :

$$\begin{cases} n_A \sim n_T \\ n_C \sim n_G \end{cases} \tag{1.37}$$

where n_a here represents the number of nucleotides a on one strand. Although this statistical property of DNA — today known as Chargaff's second parity rule (or PR2) — suffers from some exceptions, notably in mitochondria [18], it is fulfilled in a wide range of species [17, 18]. The formal explanation for this rule has been found 20 years ago, when Lobry and Lobry [19] showed analytically that this property could be simply explained under the no-strand bias condition [20],

which states that mutations affect similarly both strands of DNA. As we have seen before, whenever a mutation occurs, nucleotides on both strand are changed. For instance, if an A is replaced by a C on one strand, then a T will be replaced by a G at the same position on the complementary strand. It follows that under this no-strand bias condition, the mutation rate associated to these two mutations has to be the same. Similarly, each of the twelve possible mutations has one equivalent mutation, and thus 6 mutation rates are enough to model the evolution of DNA, such that the instantaneous rate matrix is given by:

$$Q = \begin{pmatrix} \bullet & \mu_{GT} & \mu_{CT} & \mu_{AT} \\ \mu_{AC} & \bullet & \mu_{CG} & \mu_{AG} \\ \mu_{AG} & \mu_{CG} & \bullet & \mu_{AC} \\ \mu_{AT} & \mu_{CT} & \mu_{GT} & \bullet \end{pmatrix} \tag{1.38}$$

where $\bullet$ is once again defined such that the sum over each column is equal to zero. One can show analytically that the evolution of a sequence according to a Markov model with such an instantaneous rate matrix reaches a stationary state where Chargaff's second parity rule always holds [19]. The fact that this rule is fulfilled in the genome of the majority of species indicates that most of the time, genomes evolve according to the no-strand bias condition.

Unlike the first parity rule however, the second parity rule is not exact. Although this rule is most of the times fulfilled on the global scale (when a large portion of genome is considered), deviations have been found at the local scale, indicating that in some regions of the genome, the mutations do not affect both strand symmetrically [21–24]. Studying the deviation from PR2 has revealed itself a powerful tool to highlight a wide amount of features of specific regions [25], as for instance the position of replication origins [26].

Chargaff's second parity rule gives a good example of the two major interests of simple models in biology. First, while understanding simple statistical features, one can get insight into global and basic properties of biological processes.

Secondly, they offer a global framework from which one can identify local and peculiar deviations. Analyzing and explaining these deviations can help to identify new phenomena and to develop a refined view of biological processes.

1.2.1.4 Match Length Distributions

The goal of this thesis is to study biological processes that generate long repeated segments in DNA sequences, and that are not taken into account in the models we have presented so far. To study these biological processes, we focus on the distribution $M(.)$ of the length of exact matches (segments with an identical sequence) which are maximal (i.e. they cannot be extended on either side). Such distributions can be obtained for either a self-alignment (by aligning a genome against itself), or for a comparative alignment (by aligning two different genomes without allowing for gaps and mismatches).

For sequences generated with the iid model, this distribution is given by an geometric distribution:

$$M_{\text{iid}}(r) = \frac{1}{2} L_1 L_2 (1-p)^2 p^r, \tag{1.39}$$

where L_1 and L_2 are the length of the two sequences and p is the probability that two nucleotides match, namely:

$$p = f_A^{(1)} f_A^{(2)} + f_C^{(1)} f_C^{(2)} + f_T^{(1)} f_T^{(2)} + f_G^{(1)} f_G^{(2)}, \tag{1.40}$$

where $f_a^{(i)}$ represents the frequency of the nucleotide a in the sequence i. Note that this geometric distribution leads to the well-known Gumbel distribution for *longest* matches in an alignment of iid sequences, which is commonly used to assess the significance of local alignments [27, 28].

In eukaryotic genomes one observes large deviation from this theoretical distribution, either when comparing different genomes, or when computing the self-alignment of a genome. Namely, one finds more matches of long length than would be expected under the iid model. Interestingly, these deviations exhibit a power-law tail, that is $M(r) = \mathscr{C} r^{\alpha}$, where α is the exponent of the distribution,

and $\mathscr{C}$ a normalization constant. Therefore the goal of this thesis is to understand mathematically these power-law behaviors, and to show that different duplication mechanisms have left various footprints in genomes.

1.2.2 Complex Processes of Genome Evolution

The underlying hypotheses of the models we have discussed up to this point are quite simple and take into account the very basic features of DNA. However, many more complex evolutionary processes have been identified. The importance of these different processes on the evolution of species has been well established and is, in many cases, not fully understood. In the next section, we review some of these processes.

1.2.2.1 Transposable Elements

Transposable elements (TE), which are also called transposons or repetitive elements, consist in small DNA sequences, ranging from a few hundred bps to several kbps, and which have the ability to duplicate themselves in genomes. The content of eukaryotic genomes in transposable elements is highly variable. For instance, TE cover roughly 50% of the Human genome [29], 85% of the maize genome [30] but only 14% of the genome of *Arabidopsis thaliana* [31]. Duplications of transposons are thought to occur in short bursts, which last until the host organism finds a way to repress their duplication.

After such a burst, the existing sequences remain in the genome and neutrally fade away into the genomic background due to mutations. There exists several families of transposons, which are classified according to their duplication mechanisms and their sequence. The most common type of TE in the Human genome is the Alu element [32, 33], a roughly 300 bps TE which has duplicated more than a million times in the Human genome (for reviews about Alu and other retroelements, see Batzer and Deininger [34] and Deininger and Batzer [35]).

TE are most of the time regarded as selfish DNA elements invading the genome of their host species, and their insertion has often been associated to deleterious effects. For instance, if a transposon inserts into a gene and disrupts its sequence, it can often result in a loss of function. It has also been found that they strongly contribute to genomic instability [36].

However, evidences of their positive contribution to the evolution of species have also been found. For instance, almost 25% of human promoter regions (whose role is to recruit a RNA polymerase, and thus initiate the expression of a gene) contain sequences derived from TE [37, 38]. Similarly, in *Arabidopsis thaliana*, 7.8% of expressed genes contain a region with close similarity to a known TE sequence [39]. Bursts of transposable elements have also been associated to speciation events (as a mass insertion event occurred during the formation of primates [40]), and to adaptive evolution, notably because significant changes in the rate of TE transpositions have been identified following biotic and abiotic stress conditions in plants, and, to a lesser extent in Drosophila [37]. However, direct evidences of adaptative effects of TE bursts have not yet been found.

Interestingly, the sequence and the mode of duplication of TE is highly variable from a species to another. For this reason, it is likely that a high number of these elements have not been identified yet, and that the sequencing of new species, or the improvement of the quality of genomes which have already been sequenced, will allow to get more insight into the evolutionary history of these elements, and into their impact on the evolution of genomes.

1.2.2.2 Low Copy Number Repeats

Gene Duplication — While transposable elements are short sequences that can be found an immense number of times in genomes, the latter also exhibit sequences duplicated a few number of times. Unlike transposable elements, these low copy number repeats do not possess the ability to duplicate themselves, and result from errors that occurred during different processes of genomic evolution. Two

homologous sequences that result from a duplication event are named "paralogs" in contrast to orthologous sequences which result from the divergence of two species.

The importance of these low copy number repeats (LCNRs) on the evolution of species, which first appeared in Susumu Ohno's influential book *Evolution by Gene Duplication* [41] has been widely recognized ever since. In this book, Ohno presented the idea that a single copy of a gene is enough to fulfill a function, and thus, when a gene duplication occurs, the constraints that prevented genes to evolve would apply to one of the two copies only, leaving the other one free to mutate. While most of the time these mutations would result in the loss of function of the gene (an event called pseudogenization), the "free" gene might also sometimes gain a new function. The fate of genes following duplication is a complex matter that remains under investigation and their evolution is thought to result into one of the three following scenarios [42, 43]:

– The first possible scenario is pseudogenization. In this case, one of the two copies accumulates degenerative mutations and gets silenced.

– In a second scenario (neofunctionalization), one of the two copies gains a new advantageous function, while the other copy retains the original function.

– The last possible scenario is the subfunctionalization of the two genes. In this scenario, both copies become partially inactivated due to mutations. Both copies accumulate mutations up to the point where the conjugated function of the two duplicates is equivalent to the one of the ancestral gene before the duplication event. A special case of subfunctionalization can occur if the ancestral gene is performing several functions. In that case, each copy can specialize and improve one of the several functions at the expense of the other function (that is maintained by the second copy).

In the following, we present 3 different types of LCNRs that will be studied in this thesis.

Segmental Duplications — Segmental duplications (SDs) are usually defined as long (> 1kbp) segments of duplicated DNA sharing a high sequence similarity [44]. SDs cover a significant proportion of eukaryotic genomes, and notably 5.5% of the Human genome and 5% of the Mouse genome [44, 45]. SDs can occur through various biological mechanisms and are most of the time the consequence of aberrant repair following double strand breaks. One suggested mechanism involves the recombination of non homologous segments during meiosis, and is mediated by pre-existing long repeated segments (either Alu elements [46, 47] or pre-existing SDs), and the probability that such an event occurs has been shown to increase with the similarity between the repeats mediating the event. Another mechanism is known as non homologous end joining (NHEJ) [48, 49], which mostly occurs in subtelomeric regions. Other mechanisms generating SDs have been identified, such as fork stalling and template switching (FoSTeS) [63]. For a detailed review of the different mechanisms leading to the formation of SDs, see Hastings, Lupski, Rosenberg, and Ira [50].

As SDs are mostly side effects of errors in DNA repair mechanisms, they are not randomly distributed in genomes, and exhibit a higher frequency in regions of known instability. For instance, centromeres and, to a lesser extent, close to telomeres, are significantly enriched in SDs [49]. Notably, 31% of all duplicated bases of the Human genome are located in the 5Mb regions surrounding centromeres, resulting in a $6-7$ fold enrichment compared to other regions of the genome [51].

Interestingly, although the fate of duplicated genes is most of the time discussed, mechanisms generating segmental duplications do not target specifically genes. Only a small correlation has been found between the frequency of SDs and the gene content, but it is most likely a consequence of a higher rate of retention for SDs containing genes than of a higher rate of duplications in gene rich regions.

As SDs occur through processes involving mostly DNA, we will refer to these LCNR as DNA-mediated duplications.

Retroduplications — Segmental duplication is not the only biological process that produces LCNR in eukaryotic genomes. Retroduplication is another well known biological mechanism which consists in the reverse-transcription of a mature mRNA molecule into DNA which is then reintegrated in the genome. The duplicated sequences generated by reverse-transcription exhibit different features compared to SDs.

First, most genes are composed of a mosaic of coding sequences (exons) and non-coding sequences (introns). Hence, after the transcription of the entire gene into an mRNA molecule, its introns are spliced to produce a mature mRNA, which is much shorter than the full gene sequence. Mature mRNAs which are reverse-transcribed in genomes thus result in partial duplicates, which are on average shorter than duplicates produced by DNA-mediated mechanisms.

Second, retroduplicants consist in exonic sequences only, and thus do not contain regulatory elements and promoters that would allow them to be transcribed. For this reason, they mostly produce non-coding copies highly similar to the gene transcript, commonly known as processed pseudogenes [52, 53]. Note however that functions have been found for such processed pseudogenes, and the debate about their potential role is still open, see for instance Kaessmann, Vinckenbosch, and Long [53] or Okamura and Nakai [54]. Still, it seems that most of the time, they result in non-functional evolutionary dead-ends.

Third, only genes can be reverse-transcribed, while all sequences can be duplicated through SDs. Moreover, it was shown that the probability that a gene gets retroduplicated is proportional to its expression level (i.e. to the number of mRNA molecules to which it gives rise) [55, 56].

The consequence of these three main differences is that unlike SDs, where the two duplicates are more or less equivalent after the duplication event, the original gene and its retroduplicate copy have different properties and can most of the time be told apart. Notably, while duplicates produced by DNA-mediated mechanisms can potentially duplicate in turn, retroduplicates will not give rise to retroduplicates because they are not transcribed.

Whole Genome Duplication — The two first processes we described are thought to occur relatively frequently and to affect a small number of nucleotides. But genomes are sometimes subjected to large scale duplication events. These large scale events can lead to the duplication of a complete chromosome, or even to whole genome duplications (WGD).

Of course, these large scale events have more dramatic effects on the genome they affect. When they occur, they create a large quantity of genetic material free to evolve, and have for this reason been postulated to be an important mechanism permitting genomic innovation. They might be responsible of many speciation events [57]; for instance they are associated to 15% (resp. 31%) of such events in flowering plants (resp. in ferns) [58]. It has also been noticed that most of the time, these events are followed by a large number of deletions [59]. This last property makes these events difficult to study and to evidence. While it has been postulated more than 40 years ago that two rounds of whole genome duplications took place at the base of vertebrate lineage, this hypothesis is still debated.

1.2.3 Thesis Outline

The goal of this thesis is to study the lentgh distribution of repeated words in eukaryotic genomes. Chapter 2 lists and details the experimental procedures and mathematical models used in subsequent chapters. It can be read section by section, depending on the requirements, while reading the other chapters. In Chapter 3 we study the length distributions in the case of genome self-alignments. We first explain the mathematical framework that will be used all along the thesis, apply it to several duplication mechanisms, and discuss some deviations from our model observed in two primate genomes. In Chapter 4 we investigate the case of genome comparison and show that in this experiment, the deviations from the random model are generated by a distribution of mutation rates along different genomes. In Chapter 5, we analyze the distribution obtained in the self-alignment of genomes after a whole genome duplication event, that necessitates ingredients from both self-alignments and comparative alignments. Finally in Chapter 6, we

study the comparison of coding sequences only, and show that the simple processes introduced in the previous chapters are not enough to explain the deviations observed for these regions.

Most of the results presented in this thesis have been published in peer reviewed journals. The simplest model that only takes into account random segmental duplications (Section 3.2) has been published in *Physical Review Letters* [60]. The generalization to Yule tree like processes (Section 3.3), to processed pseudogenes (Section 3.4) as well as to comparative alignments (Section 4.4.1) appeared in *Molecular Biology and Evolution* [61]. Finally the detailed calculation of the number of leaves at a given distance on a Yule tree is the subject of an article in *Plos One* [62]. Since some of the results of this last paper are beyond the scope of this thesis, it is attached as an appendix at the end of this document (Appendix C).

Chapter 2

Materials and Methods

In this chapter, we introduce tools and concepts related to genome analyses and modeling that were defined and developed in previously by others and that will be used in the course of this thesis.

2.1 Computing MLDs

To find all identical matches (both in the case of self and comparative alignments), we used the `mummer` pipeline [64] (version 3.0), which allows to find all maximal exact matches between a "query" and a reference sequence using a computationally efficient suffix tree approach. For our analyses, we used the following options:

• `-maxmatch` such that `mummer` searches for all matches regardless of their uniqueness (by default, only matches unique in the query sequence are retrieved).

• `-n` that states that only "A", "T", "C" and "G" can match (any other character always results in a mismatch).

• `-b` such that `mummer` searches for matches on both strands. To do so, the reverse complement sequence of the query file is computed, and `mummer` searches matches between the two forward sequences, as well as matches between the forward sequence of the reference and the reverse complement sequence of the query.

• `-l 20` to filter out matches smaller than 20 (default value). Most of the time we used the default value of 20.

`Mummer`'s output consists in a file with three columns representing for each match its position in the query sequence, its position in the reference sequence and its length. The number of matches expected for a random iid sequence grows quadratically with L ($3.5\,10^{18}$ matches of length 2 when one compares two sequences of length $L = 10^9$ bps, see Eq.(1.39)), which explains why one has to define a minimum match length in `mummer` when comparing entire eukaryotic genomes. Depending on the length of the sequences to compare, we might vary the value of this threshold.

Computing a MLD from a `mummer` output is then straightforward. One needs to count the number of matches obtained for each length. To make the MLD computation available to anyone interested, we developed an online tool integrated in Galaxy. Our tool takes as input a sequence to self-align (or two sequences to compare if one wants to compute a comparative alignment), and gives as an output the MLD represented using a logarithmic binning. The user can choose the threshold for the minimum match length of `mummer`, as well as the value of the multiplicative factor of the logarithmic binning (see Section). In order to compute the MLD from length one, we slightly modified `mummer` source code so that it outputs directly the MLD. We used this version only in few cases.

2.2 Power-Law Distributions

General Properties — Power-law distributions are distributions of the form

$$p(x) = \mathscr{C} x^{\alpha} \tag{2.1}$$

where α, the exponent of the distribution, is negative and where $\mathscr{C}$ is a normalizing constant such that $\int p(x)dx = 1$. The value of $\mathscr{C}$ thus depends on the value of α. As $p(x)$ obviously diverges in $x = 0$, there must be a lower bound $x_{\min}$ from which the power-law behavior starts. This explains why power-laws often appear in the

tail of distributions, i.e. for values larger than a certain lower bound. If x takes values from $x_{\min}$ to infinity, it follows that

$$\mathscr{C} = \left(\int_{x_{\min}}^{\infty} x^{\alpha} \, dx \right)^{-1} = -\frac{\alpha + 1}{(x_{\min})^{\alpha - 1}} \ , \tag{2.2}$$

when $\alpha < -1$. Distributions with $\alpha \geq -1$ are not normalizable, and need to be bounded by a maximal value. The distributions we study in this thesis all have an exponent $\alpha < -1$, so we disregard this case in the following and always assume that $\alpha < -1$.

Several features of these distributions can explain why they have attracted a wide interest in the scientific community. Notably, these distributions have a "heavy tail", meaning that the probability of extreme events is much higher than for other classical probability distributions, such as the normal distribution. Another interesting property of power-laws is that these distributions do not depend on the scale one looks at it. Notably, one can establish that

$$p(bx) = g(b)p(x) \ \forall x, \tag{2.3}$$

meaning that if we "zoom in" or "zoom out" (for instance by changing the unit in which the distribution is measured), the shape of the distribution stays unchanged. Let us explain this with a small example. Let $p(x)$ stands for the number of scientific papers which have been cited x times (a quantity that has been shown to be power-law distributed [65]), and suppose that we find that there are 10 times more articles that have been cited twice than papers that have been cited 4 times (i.e. $g(2) = 1/10$). Then, we would also find that there are 10 times more articles that have been cited 20 times than papers that have been cited 40 times. Interestingly, power-law distributions are the only distributions to fulfill the relationship defined in Eq. (2.3). For this reason power-laws are said to be "scale free" or "scale invariant" distributions.

A tremendous amount of phenomenon have been reported to follow power-law like distributions, such as the frequency of words in human languages, the frequency

of use of names in most cultures, the number of species in a biological taxa or the magnitude of earthquakes to cite only a few examples (for a review on the topic, see Newman [66]). In genomic studies as well, many power-law distributions have been identified [67–69]. As the number of phenomenon found to follow power-law distributions was growing, attempts to unify all different processes under a few mechanistic models, and to explain why such distributions were so common have emerged [70–72].

As a consequence, questions regarding the inherent interest of these distributions, together with the claim that many reported power-laws were lacking good statistical support have also emerged [73, 74], together with the development of appropriate mathematical frameworks to study these distributions [74].

The goal of this thesis is not to participate in this wide and passionated debate. We can note however that in many cases, valuable insight to the understanding of biological processes, for instance to genome evolution, has been gained from the understanding of these power-laws such that it remains worth studying [72, 75–77].

Representing Power-Laws — To illustrate the properties of power-law distributions, we show a synthetically generated dataset following a power-law distribution on Fig. 2.1(a). This figure was reproduced from Newman [66] (see this article for detail explaining how to synthetize such a set). As simulated data are continuous, to compute an histogram, one has to generate bins and to count the number of data points observed in each of these bins. The histograms of Fig. 2.1(a) and (b) are generated using a constant bin size.

Taking the logarithm of both side of Eq. (2.1), one finds that:

$$\ln\left(p(x)\right) = \alpha \ln(x) + \ln(\mathscr{C}), \tag{2.4}$$

and it follows that, represented on a log-log scale, a power-law distribution appears as a straight line, whose slope is equal to α. This behavior can be observed on Fig. 2.1 panel (b).

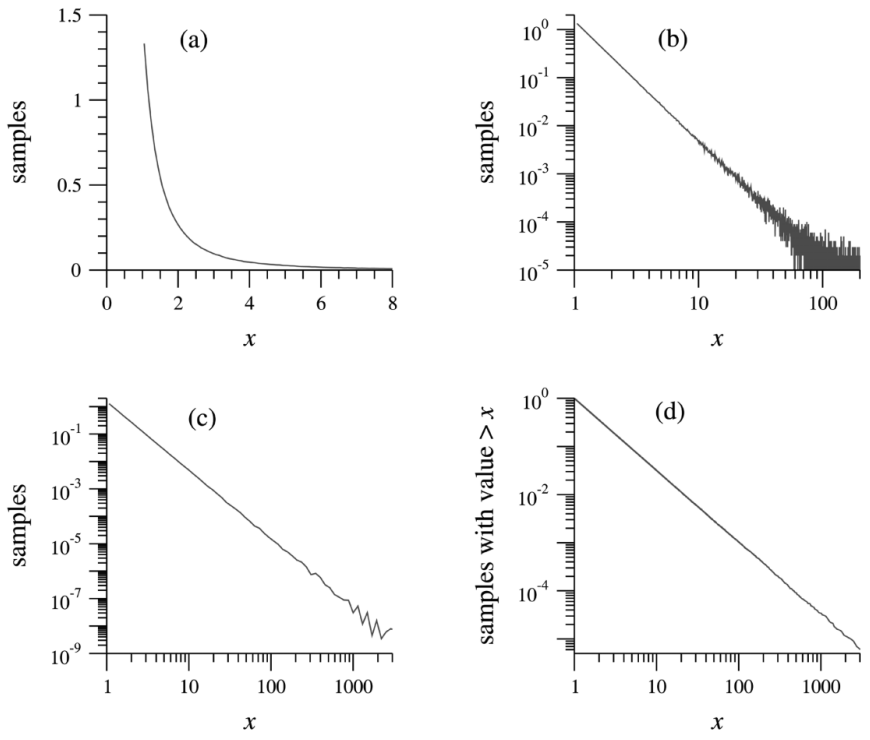

FIGURE 2.1: In this figure, we show four different representations of the same data. Data were synthetically generated to follow a power-law distribution with an exponent $\alpha = -2.5$. (a) Representation on a linear scale (b) Representation on a log-log scale (c) Representation on a log log scale using logarithmic binning (d) The cumulative distribution. Figure reproduced from Newman [66]

This property makes it really convenient to detect power-law distributions. Note however that one should not only rely on the impression that raw data is a straight line on a log-log plot to conclude that data are power-law distributed [73, 74]. Similarly, it has been noted that evaluating the exponent of the power-law by fitting a straight line to the data on a log-log plot was a biased method [66, 78].

Logarithmic Binning — Power-laws appear in the tail of distributions, meaning that they are associated to rare events, which are thus subject to strong variations. The high impact of noise in the tail of the distribution – making the assessment of the exponent of the distribution difficult – can be observed on the synthetic set represented on Fig.2.1 (b). A way to resolve this issue is to increase the size of the bins with the value of x. Of course, such procedure introduces a bias in the distribution, and one then has to normalize the data. Namely, the value observed for each bin is divided by the size of the bin. The most common choice to do

this is known as the logarithmic binning procedure, which consists in increasing the size of the bin by a constant factor. For instance, if the first bin is of size 0.2 and the multiplying factor is set to 2, then the second bin will be of size 0.4, the third of size 0.8, the fourth of size 1.6, and so on. Note that by doing this, one dramatically reduces the number of data points. Such a procedure was used to generate Fig. 2.1 panel (c). On this figure, one can see that it nicely reduces the noise, and makes the power-law behavior much clearer. Note that binning induces a loss of information, as one aggregates different data points together in the same bin (as data with different values are summarized together as one data point), so that it is often useful to consider both representations.

Another procedure to reduce the noise in the tail of the distribution consists in representing the cumulative distribution $P(.)$, which is defined as:

$$\begin{aligned} P(x) &= \int_x^\infty p(a)\,da = \int_x^\infty \mathscr{C} a^\alpha\,da \\ &= \frac{\mathscr{C}}{\alpha-1} x^{\alpha+1}. \end{aligned} \tag{2.5}$$

The cumulative distribution also follows a power-law distribution (although with an exponent $\alpha+1$). As can be seen on Fig. 2.1 panel (d), this procedure also efficiently reduces the noise, and is often reported as superior to the logarithmic binning (notably by Newman [66]). However, in the case of discrete power-law distribution, logarithmic binning still seems more efficient [79]. As we study discrete distributions in this thesis, we chose to always represent data using the logarithmic binning procedure.

Once the power-law behavior is established, a method to obtain an estimate of the value of α, is to compute the maximum likelihood estimator. The estimator $\hat{\alpha}$ is then the value of α that maximizes the log likelihood $\mathcal{L}$:

$$\mathcal{L} = \sum_{i=1}^{n} \left[\ln(\alpha-1) - \ln(x_{min}) - \alpha \ln\left(\frac{x_i}{x_{min}}\right) \right], \tag{2.6}$$

such that

$$\hat{\alpha} = -1 - n \left[\sum_{i=1}^{n} \ln \left(\frac{x_i}{x_{\min}} \right) \right]^{-1}, \tag{2.7}$$

while the value of $x_{\min}$ has to be determined visually. This estimator is also sometimes referred to as the Hill estimator [80].

The power-law distributions that we analyze in this thesis span roughly one order of magnitude (from length 20 to 200 roughly). Notably, for most of the distributions we study, the very tail of the distribution exhibits a faster decrease than expected for a power-law. Such behavior are sometimes modeled with an exponential cutoff, that is:

$$p(x) = \mathscr{C} x^{\alpha} \exp(-\lambda x). \tag{2.8}$$

The MLE, that does not take this cutoff into account, is a biased estimator of the value of the exponent.

More generally, the goal of the thesis is to try to understand large deviations of match length distributions from the random iid model (see Section 1.2.1.4). It might be that more complex distributions, comprising features such as an exponential cutoff, more accurately describe our data. Our aim is to develop mechanistic models that result in deviations similar to those observed in our data, and that are biologically meaningful, not to claim that the data we observe are power-law distributed. As power-law distributions seem to be fair approximations of the distributions we observe, we will in the following, for simplicity (and maybe sometimes abusively) describe them as power-law distributions.

2.3 Yule Trees

A Yule tree [81] (also known as a birth death tree) is the result of a branching process with constant birth and death rate. As one of the simplest stochastic models for branching processes, it is often used to model the evolution of families of species or of families of genes that evolve from a common ancestor. A Yule tree is defined as follow. At the beginning of the process (at $t = 0$), the tree consists of

only one individual. During each infinitesimal time interval dt, this individual can either give birth to a new individual (with probability λdt) or die (with probability δdt). λ and δ are defined as the birth and death rate of the process respectively. We show an example of such a tree on Fig. 2.2.

Depending on T, the total time elapsed from the beginning of the process (also called the height or the age of the tree), one can calculate several useful quantities, that we will use in Section 3.3. Let $P(Z|T)$ be the probability that there are Z leaves on a tree of age T. Following [82], as long as the birth rate is larger than the death rate, the average number of leaves in a Yule tree grows exponentially with the age of the tree and is simply given by:

$$E(Z|T) = \exp\left[(\lambda - \delta)T\right]. \tag{2.9}$$

We can also write the probability that no individual ($Z = 0$) survives through to time T, that is:

$$P(Z = 0|T) = 1 - \frac{\lambda - \delta}{\lambda - \delta e^{(\delta-\lambda)T}}. \tag{2.10}$$

Finally, for $z > 0$ we have [82]:

$$P(Z = z|T) = \frac{\lambda - \delta}{\lambda - \delta e^{-(\lambda-\delta)T}} \left[1 - \frac{1 - e^{-(\lambda-\delta)T}}{1 - \frac{\delta}{\lambda} e^{-(\lambda-\delta)T}}\right] \left[\frac{1 - e^{-(\lambda-\delta)T}}{1 - \frac{\delta}{\lambda} e^{-(\lambda-\delta)T}}\right]^{z-1}. \tag{2.11}$$

2.4 Simulating the Evolution of DNA Sequences

To simulate our evolutionary models, we used the following process. A sequence of nucleotides $\mathcal{S} = (s_1, \ldots, s_L)$ of length L with $s_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ is evolved through time using small time intervals Δt. Time intervals Δt are small enough such that for any evolutionary process e of the model occurring with rate ρ_e, we have $\rho_e L \Delta t \ll 1$. At each step, a random number (u_i^e per position i and per possible evolutionary process e is drawn from a uniform distribution. The event e then

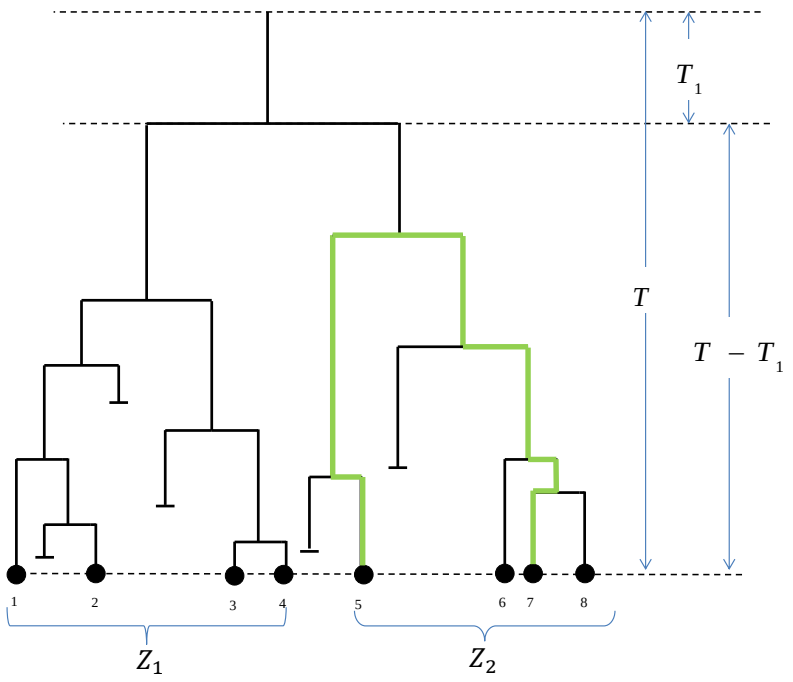

FIGURE 2.2: An example of a rooted Yule tree of height T with 5 leaves. The pairwise evolutionary distance between two leaves (green path) is denoted by τ. The first branching event occurs after time T_1 and the two resulting subtrees possess Z_1 and Z_2 leaves. Small horizontal lines represent dead leaves. The horizontal dimension is meaningless.

occurs at position i if $u_i^e < \rho_e \Delta t$. These steps are repeated until the desired time t has elapsed.

2.5 Bioinformatic Procedures

Genomes — Unless otherwise stated, all genomes and their specific annotations (such as repeated elements and exons) were downloaded from the `ensembl` website [83]. For Human, we use GRCh37 release. *Arabidopsis thaliana* is the only species we studied whose genome is not available on the `ensembl` website, and we downloaded it from the `TAIR` website [84].

Phylogenetic Tree of Pseudogenes — As an example of Human processed pseudogenes, we analyzed in Section 3.4 the behavior of the familly of processed pseudogenes that results from the retroduplications of the RPL21 gene, a gene that retroduplicated many times in the human genome. To define the subset of RPL21

pseudogenes, we searched the set of all human pseudogenes (downloaded from the `pseudogene` database), for all sequences homologous to the RPL21 transcript using `BLAST` algorithm [85]. We kept only the sequences with an alignment score larger than half of the length of the RPL21 transcript. This resulted in 117 sequences. We computed the multiple alignment of all these sequences using the `MAFFT` program [86] in the most accurate mode (`linsi`). Afterwards, we cleaned the alignments to keep only the most reliable positions of the alignment with the `trimAl` program [87] in automatic mode. To calculate the distance matrix summarizing all the pairwise distances between the different pseudogenes, we used the package `phylip` [88]. Four sequences were excluded due to their large distances to other sequences. After calculating the distances, all pseudogenes were ranked according to their average distance to other pseudogenes, from small to large. Then we assumed that the topology of the phylogenetic tree is such that the gene is retroduplicated to the first pseudogene in the ranking and then to the second one, etc., resulting in a ladder tree, as shown in the next Chapter on Fig. 3.10. The tree was built using the same `phylip` package but with a fixed topology (that is the one discussed in the previous sentence). For this procedure, we used the F84 model [89] for nucleotides substitutions.

RepeatMasker — The RepeatMasker [90] pipeline is a tool that screens DNA sequences to identify repetitive elements and low complexity DNA sequences (which are simple repeated sequences, such as "GTGTGTGTGTGTGTGT..." for instance). Using the Smith-Waterman-Gotoh algorithm [91, 92], it searches the query DNA sequences for regions that share a high similarity to previously identified repetitive elements, which are listed in the RepBase database [93]. It then produces a "RepeatMasked" sequence, which is similar to the query sequence, but where all letters that have been identified to be part of a repetitive element have been replaced by an "N". RepeatMasked version of eukaryotic genomes can also be directly downloaded from the `ensembl` database [83].

Chapter 3

Self-alignment

In this Chapter, we explain the different power-laws observed in the Match Length Distributions of the self-alignment of eukaryotic genomes, and demonstrate why different evolutionary scenarios give rise to power-laws with different exponents.

We first discuss the general mathematical framework that we will use in the thesis (Section 3.1) , and then apply this framework to three evolutionary scenarios. In the first and simplest scenario (Section 3.2), we consider segmental duplications as a random and continuous process. We then extend our result to two more complex and more biologically relevant scenarios. In the first of these two, a particular sequence segment and its duplicated offspring duplicate again and again, giving rise to a family of duplicated genes (Section 3.3). The second scenario is meant to represent the duplication through the reverse-transcription of the mRNA of a gene, giving rise to many pseudogenes (Section 3.4).

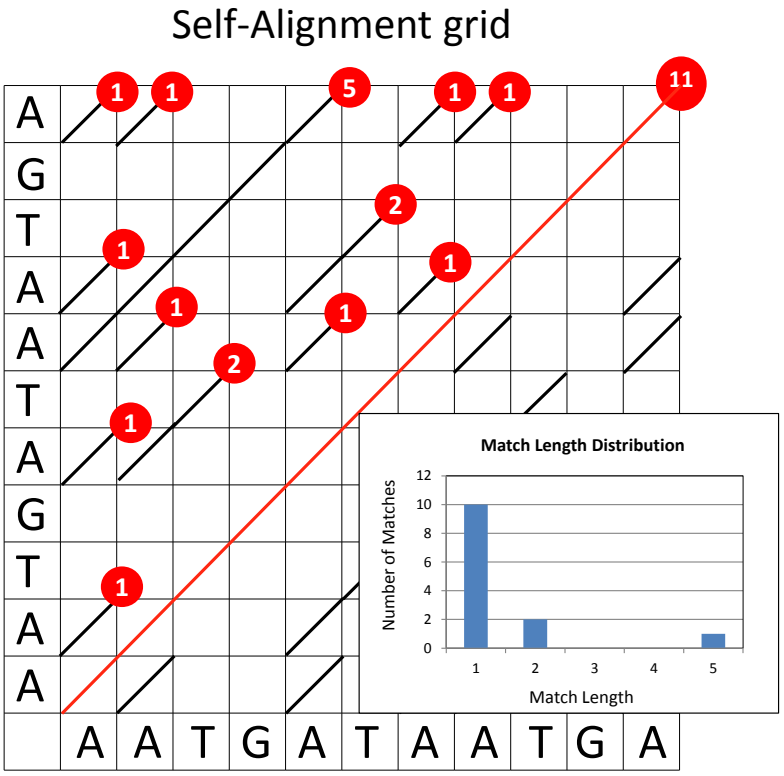

FIGURE 3.1: A toy example of the self-alignment of a small sequence and its corresponding MLD. Black lines in the grid represent exact matches and values in the red circles represent the length of maximal exact matches in number of base pairs.

3.1 Preliminary Considerations

3.1.1 The Match Length Distribution of the Self-Alignment of a Genome

The goal of this thesis is to explain a set of puzzling statistical properties of eukaryotic genomes that was first described by Gao and Miller [94] (see also Sindi [95], Csűrös, Noé, and Kucherov [96] or Salerno, Havlak, and Miller [97] for related analysis). The authors of this article studied a statistical property of Repeat-Masked genomes. Namely, they aligned each genome against itself to retrieve all identical matching segments. We will refer to this experiment as the self-alignment of a genome in the following. For their analyses, they only took into account exact gapless matches which are maximal, in the sense that these matches cannot

be extended on either side (i.e. they are not included in longer matches). They then counted the number of matches observed for each length to obtain the Match Length Distribution (MLD), see Fig 3.1 for a simple example of the procedure on an 11 bps sequence. We reproduce this experiment for the RepeatMasked human genome, and for a sequence produced from the concatenation of all human exons (also called the exome) in Fig. 3.2. To retrieve all matches, we used the `Mummer` pipeline [64], see Section 2.1 for details. Note that all maximal matches are considered here, regardless of their uniqueness.

For small matching segments of lengths $r < 20$, the MLD follows the expected distribution $m_{iid}(r) = L^2 p^r (1-p)^2$, obtained for a random iid sequence (see Section 1.2.1.4). However, for longer matches, one observes many more matches than expected: given the length of the human genome, we expect no match longer than 30 bps. Even more surprisingly, the distribution for these long matches follows a power-law distribution with exponent $\alpha = -3$.

In the following, we study the behavior of what we call the tail of the MLD, that we define as the distribution in the range where the power-law behavior is observed. The "lower bound" $r_{\min}$ of this tail is the shortest length for which the distribution exhibits a power-law behavior. $r_{\min}$ depends both on the intercept of the power-law distribution and on the number of randomly expected matches (which itself mainly depends on the total length of the analyzed sequence), and is equal to $r_{\min} = 20$ in a typical eukaryotic genome.

3.1.2 The Stick Breaking Process on Evolutionary Time Scale

In this chapter, we want to show that this power-law distribution can arise from the neutral evolution of segmental duplications. Before studying the Match Length Distribution at the genomic scale, let us focus on the evolution of a single segmental duplication. Just after its generation, a segmental duplication will consist in two 100% identical sequences of length K, resulting in one match of length K. As

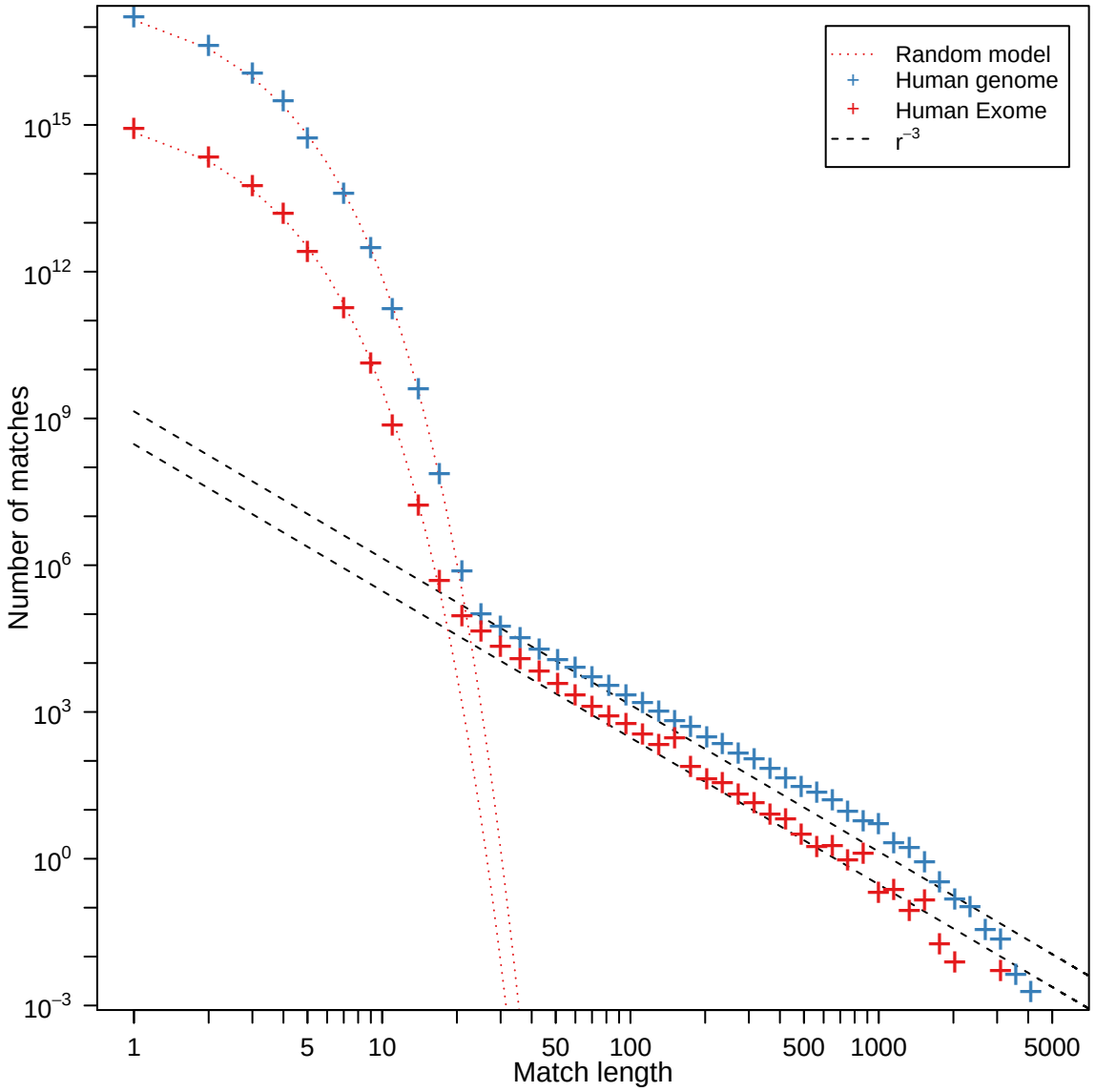

FIGURE 3.2: The match length distribution (MLD) computed from the self-alignment of the RepeatMasked human genome and the RepeatMasked human exome. The red dotted lines represent the distribution obtained when repeating the same experiment on a random iid sequence with same length and equal nucleotide frequencies. The dashed lines represent the power-law functions L/r^3 and $3L_{ex}/r^3$, where L is the length of the RepeatMasked human Genome (we do not count the Ns), L_{ex} is the length of the human exome and where r is the match length. Both MLDs are represented using logarithmic binning to reduce the sampling noise, see Section 2.2 for a discussion on this subject.

time passes, this match will be disrupted by mutations. When a match of length r is broken by a mutation, two smaller matches of length r_1 and r_2 such that $r = r_1 + r_2 + 1$. For simplicity, we assume that the stick length r is a continuous parameter and that mutations only break a stick without shortening it (i.e. $r_1 +$

$r_2 = r$). We also performed similar analyses for the discrete case, and obtained similar results, see Appendix A. Here we focus on neutrally evolving DNA. In this case, all mutations have the same probability of fixation, and thus occur randomly at any position of one of the two copies of the duplicated sequence. This fragmentation process of matching sequence segments can be mathematically described by the stick-breaking process, a process that was first introduced to understand the fragmentation of long polymer chains [98]. In this framework, a match resulting from a segmental duplication is considered to be one full length stick, which is then broken up by mutations into several smaller sticks.

Following Ziff and McGrady [99], the dynamics of the length distribution of fragmented sticks in time can be described by a differential equation. If we define τ as the evolutionary distance (also called the divergence) between the two duplicates, and $m(\cdot, \tau)$ the length distribution of sticks (or matches) for a divergence τ between the two duplicates, then it fulfills:

$$\frac{\partial m(r,\tau)}{\partial \tau} = -2\mu r\, m(r,\tau) + 4\mu \int_r^\infty m(s,\tau)\, ds. \tag{3.1}$$

The first term represents the loss of matches of length r due to mutations. These mutations occur with rate μ, and can break any of the $m(r,\tau)$ matches at any of the r positions of each of the two copies, so this event occurs with probability $-2\mu r\, m(r,\tau)$. The second term describes the gain of matches of length r resulting from a mutation in a longer match. In any match of length $s > r$, there are 4 positions (2 in each copy) where the occurrence of a mutation results in a match of length r (and a match of length $s - r$), so this event has a probability $4\mu \int_r^\infty m(s,\tau)\, ds$. At time $t = 0$ we start with one stick of length K, so we have the initial condition: $m(r, 0) = \delta(r, K)$, where $\delta(r, K)$ denotes the Kronecker delta (i.e. the function which is equal to zero everywhere except for $r = K$, where it takes the value 1). The analytical time dependent solution of this differential

equation with this initial condition is known to be [99]:

$$m(r,\tau) = \begin{cases} [2\tau + \tau^2(K-r)]\exp(-\tau r), \text{ for } 0 < r < K, \\ \exp(-K\tau)\delta(r,K), \text{ for } r = K, \\ 0 \text{ otherwise.} \end{cases} \tag{3.2}$$

For large τ and small r this is basically an exponential function in r.

Note that in real genomes, point mutation is not the only mechanism that can disrupt (or break) a match in several smaller matches. Indeed, insertion and deletions of DNA segments also break matches. Unfortunately, our framework does not allow to differentiate between these different mutational processes and to infer the relative contributions of each of these processes. For this reason we always consider the mutation rate as an effective rate subsuming effects of different biological processes. Note that we use an infinite allele model: in reality, a mutated loci can mutate a second time and return to its previous state (this process is called reverse mutation). When it occurs, a formerly broken match is reconstructed. We neglect this effect.

3.1.3 A Mathematical Framework to Calculate Match Length Distributions

From the previous section, we have an equation describing the evolutionary fate of a single segmental duplication. In a eukaryotic genome there are many such duplications. In this section, we develop a mathematical framework to describe the MLD resulting of many segmental duplications.

When a duplication occurs, it generates two identical DNA segments which then evolve independently from each other. Then, any of the two duplicated sequences can duplicate again, giving rise to a family of sequences. The evolution of such a family can be well described by a branching process, giving rise to a phylogenetic tree. On such a tree, leaves represent paralogous DNA segments that share a common ancestor.

Any two leaves of the tree are separated by an evolutionary distance τ from each other. This distance depends on the real time since the duplication event, and on the mutation rate along all branches separating the two duplicates. Thus, the dimensionless evolutionary time (or divergence) separating a pair of leaves is defined as

$$\tau = \sum_i \mu_i T_i, \tag{3.3}$$

where the sum runs over all the branches along the evolutionary path between the two leaves and T_i and μ_i are the length (in real time) and the mutation rate of branch i respectively. Recall that by mutation, we mean any event that disrupts a match (including single nucleotide substitutions, short insertions and deletions). For any tree j, we can define $N_j(\tau)$ the number of pairs of duplicated segments separated by an evolutionary time τ. After the first duplication event, the tree has only two leaves, and $N_j(\tau)$ is the simple function:

$$N_j(\tau) = \begin{cases} 1 & \text{for } \tau = \tau_d, \\ 0 & \text{otherwise,} \end{cases} \tag{3.4}$$

where τ_d is the divergence between the two duplicates. For larger trees, the functional form of $N_j(\tau)$ depends on τ and on the topology of the tree, as we will discuss in Sections 3.3 and 3.4.

A genome is composed of many active duplicating segments, that gives rise to many families, each represented by one tree. The total value of $N(\tau)$ for a genome is then defined as the sum of all $N_j(\tau)$ such that:

$$N(\tau) = \sum_j N_j(\tau). \tag{3.5}$$

Furthermore, from the previous subsection, we have a theoretical formula for the number of identical sequence matches of length r for a pair of sequences of length K separated by an evolutionary distance τ, see Eq. (3.2). Thus, the match length distribution from the self-alignment of a genome, M, can be simply obtained by

integrating over all pairs of duplicated segments:

$$M(r) = \int_0^\infty m(r,\tau)\, N(\tau)\, d\tau. \tag{3.6}$$

In the following, we study the value of $N(\tau)$, and the resulting MLD for different evolutionary scenarios.

3.2 The Simplest Case: Random Duplications

3.2.1 Theoretical Calculations

Before studying more general scenarios, we start by focusing on a simplified scenario, and set the three following assumptions.

– (H1): We assume that duplications occur uniformly in space, i.e. that duplications have the same probability to occur at any position of the genome.

– (H2): We further assume that all duplicated segments have the same length K, and that K is negligible compared to the length of the studied genome L, meaning that duplications do not change the size of the genome. The combination of these first two assumptions yield that the probability that a given region is duplicated more than once is negligible.

– (H3): Finally, we assume that the duplication rate is constant (i.e. there are no burst of segmental duplication and no long period without any such event).

Taken together, these three assumptions lead to the fact that on average, for any given evolutionary time τ, there is a constant number of pairs of duplicates separated by an evolutionary distance τ from each other, i.e.:

$$N(\tau) = N_0, \quad \forall\ \tau \tag{3.7}$$

where N_0 is a constant.

Replacing the value of $N(\tau)$ in Eq. (3.6), we obtain that $M(r)$, the number of exact matches of length r obtained in the self-alignment of a genome is equal to:

$$M(r) = \int_0^\infty N_0 m(r,\tau)\, d\tau. \tag{3.8}$$

Replacing $m(r,\tau)$ by its value given in Eq. (3.2), we obtain for $0 < r < K$:

$$M(r) = N_0 \int_0^\infty 2\tau \exp(-\tau r)\, d\tau + N_0 \int_0^\infty \tau^2 (K-r) \exp(-\tau r)\, d\tau. \tag{3.9}$$

Integrating by parts on the second term gives:

$$\begin{aligned} M(r) &= N_0 \int_0^\infty 2\tau \exp(-\tau r)\, d\tau + N_0 \left[0 + \frac{2(K-r)}{r} \int_0^\infty \tau \exp(-\tau r)\, d\tau\right] \\ &= N_0 \frac{2K}{r} \int_0^\infty \tau \exp(-\tau r)\, d\tau. \end{aligned} \tag{3.10}$$

Integrating by parts again leads to:

$$M(r) = \frac{2N_0 K}{r} \left[0 + \int_0^\infty \frac{\exp(-\tau r)}{r}\, d\tau\right]. \tag{3.11}$$

And finally, we get:

$$M(r) = \frac{2KN_0}{r^3} \tag{3.12}$$

for $0 < r < K$, which is a power-law with an exponent $\alpha = -3$, as observed in the MLD of real genomes. For $r = K$, we simply get:

$$\begin{aligned} M(K) &= \int_0^\infty \exp(-\tau K) d\tau \\ &= \frac{N_0}{K}. \end{aligned} \tag{3.13}$$

Note that this function exhibits a non-continuous peaks in $r = K$. We observed this peak in simulated data (data not shown). In real genomes however, it is expected that the length of segmental duplications (SDs) is not constant, such that we do not observe this peak.

Thus, the power-law observed in the self-alignment of eukaryotic genomes can be

simply explained by the interplay of segmental duplications and point mutations. The appearance of a scale-invariant distribution in a process that is observed at different time points is not unexpected [100]. Surprisingly, in this integrated stick-breaking model the exponent is universal in the sense that it does not depend on the microscopic details of the model, namely the mutation and duplication rates, the length of a duplication K or the total length of the sequence L. For a more detailed discussion about the integrated version of the broken stick model, see also Ben-Naim and Krapivsky [101].

The Stationary State with Continuous Duplications — To deduce the correct normalization factor N_0 for the distribution M given by (3.12), we consider a stick-breaking process in which, according to our evolutionary model, segmental duplications of length K are continuously generated with rate λ per site. The dynamics of Eq. (3.1) for the distribution $m(r,\tau)$ then gains a third term on the right hand side which describes the influx of new matches of length K in a genome of total size L:

$$\frac{\partial M(r,\tau)}{\partial \tau} = -2\mu r\, M(r,\tau) + 4\mu \int_r^\infty M(s,\tau)\, ds + \lambda L \delta(r,K). \tag{3.14}$$

In this setting we are interested in the stationary state distribution $M_\infty(r)$ and solve the differential Eq. (3.14) for $\partial M_\infty/\partial\tau = 0$.

Let us assume that our solution is of the form:

$$M_\infty(r) = \begin{cases} N_0/r^3 & \text{for } r < K, \\ B & \text{for } r = K, \\ 0 & \text{otherwise.} \end{cases} \tag{3.15}$$

As there are no matches longer than K, for $r = K$, Eq.(3.14) becomes:

$$0 = -2\mu K\, B + \lambda L \tag{3.16}$$

and so:

$$B = \frac{\lambda L}{2\mu K}. \tag{3.17}$$

For $r < K$, we have to split the second term of the equation to calculate it:

$$4\mu \int_0^\infty M(s,\tau)\, ds = 4\mu \left(\int_0^K M(s,\tau)\, ds + \int_K^\infty M(s,\tau)\, ds \right) \tag{3.18}$$

so that the full equation results in:

$$0 = -2\frac{N_0 \mu}{r^2} + 2\mu \left(\frac{N_0}{r^2} - \frac{N_0}{K^2} \right) + 4\mu B \tag{3.19}$$

which finally leads to the solution

$$M_\infty(r) = \frac{\lambda K}{\mu} \frac{L}{r^3} \tag{3.20}$$

for $r < K$ and $M_\infty(r) = \lambda L/(2\mu K)$ for $r = K$. We can fit our observations even better by considering a discrete version of the stick-breaking model, see Appendix A. In essence these considerations yield a finite size correction to the power-law behavior in Eq. (3.20) for small r. However, we note that this correction is always negligible in the regime $r > 20$, which is the only regime where the power-law can be observed in real genomes.

From Eq. (3.20), one can also notice that the dependency of the MLD M in the length of duplicated segments, K, is purely linear. Thus, we can relax the assumption on the fixed length of the duplicated sequences with no qualitative changes on the MLD. To be able to apply our framework, as we assumed above that matches where all smaller than the length of duplicated segments K, we need the lengths of most segmental duplications to be longer than the length r_{one} where the power-law tail has value one, i.e. :

$$\left(\frac{\lambda K L}{\mu} \right)^{1/3} = r_{\text{one}}. \tag{3.21}$$

If D denotes the length of duplicated segments, this results in:

$$P(D < r_{\text{one}}) \simeq 0. \tag{3.22}$$

In that case, we retrieve the same solution for M than the one stated in Eq. (3.20), where K is now the mean length of a segmental duplication.

3.2.2 Simulations

To confirm our theoretical results, we performed simulations. We introduce a sequence evolution model that includes two basic evolutionary processes: point mutations and duplications of sequence segments. Both processes act on a sequence of nucleotides $\mathcal{S} = (s_1, \ldots, s_L)$ of L nucleotides taking values in $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}^L$.

Mutations — The mutation process induces a change in the sequence $\mathcal{S} \to \mathcal{S}'$ at one random position p, such that $\mathcal{S}' = (s'_1, \ldots, s'_L)$ is given by:

$$s'_i = \begin{cases} s' & \text{with } s' \neq s_i \text{ for } i = p, \\ s_i & \text{otherwise.} \end{cases} \tag{3.23}$$

This process happens with rate μ per site, i.e. in an infinitesimal small time interval dt it occurs with probability $\mu L\, dt$.

Duplications — The second process in our model generates segmental duplications. A random segment of $K \ll L$ consecutive nucleotides starting at a random position c in S, $(s_c, \ldots, s_{c+K-1})$, is copied and pasted to a random position v. The rest of the sequence stays unchanged; the new sequence $\mathcal{S}'$ is given by:

$$s'_i = \begin{cases} s_{i-v+c} & \text{for } i \text{ with } v \leq i < v + K, \\ s_i & \text{otherwise.} \end{cases} \tag{3.24}$$

This process overwrites the K preexisting nucleotides s_{v+k} for $0 \leq k < K$ at the target sites, and the total sequence length L stays constant, to simplify the computational procedure. This is equivalent to coupling any duplication to a deletion of a sequence of the same size. We also implemented a version where duplications are added at the end of the sequence, and do not overwrite the preexisting nucleotides. The self-alignements computed from sequences generated with the two different versions of the model exhibited similar MLDs. For simplicity we also assumed periodic boundary conditions and identify s_1 with s_{L+1}, in order to avoid any border effect. Segmental duplications occur with rate λ per site, and we assume that $\lambda \ll \mu$.

Results — Given the above processes, it is easy to simulate sequences and to perform a self-alignment to find exactly repeated segments. We start each simulation with a random iid sequence with equal nucleotide frequencies. This sequence is then subjected to the above dynamics for a time longer than t_0, which is defined as the time at which on average, each nucleotide has been mutated and duplicated more than once. After such a time, we reach a stationary state for the MLD. We then computed the MLD of the generated sequence. For more details on the simulation process, see section 2.4.

On Fig. 3.3, we show the MLDs obtained for several simulations with different values of the mutation rate μ and the duplication rate λ for sequences of length $L = 10^6$ and with a duplication length $K = 1000$. All distributions share the same behavior for small lengths ($r \lesssim 20$). That part of the distribution is dominated by small random matches which are exponentially distributed, as described in Eq. (1.39), as shown on Fig. 3.3.

For longer lengths ($r > 20$), we observe many more exact matches than expected for random sequences. As expected from the previous calculations, the length distributions of long matches follow a power-law distribution with exponent $\alpha = -3$. Moreover, we retrieve all the predictions from our theoretical calculations (see

Fig. 3.3): the distributions we obtain fit our prediction exactly, and the shape of the distribution does not depend on the values of the parameters (λ, μ, L and K).

Finally, we also simulated the case where the length of segmental duplicated segments, K, is not constant, but instead distributed according to a normal distribution with mean K and standard deviation 100. As expected, the MLD obtained in that case is equivalent to the one obtained with a fixed duplication length, if we replace the value of K of Eq. (3.20) by the mean duplication length.

3.2.3 Discussion

In this section, we introduced a simple model of genome evolution accounting for segmental duplications and mutations that gives us insights into the occurrence of a power-law tail in the length distribution of exact matches in self-alignments of genomic sequences. Using an extended version of the stick-breaking process for fragmentation, we correctly deduced the empirically observed exponent, namely, $\alpha = -3$ of this power-law distribution.

For the human genome, this tail comprises exact matching sequences of lengths varying from 25 to about 1000 bps, see Fig. 3.2. From our analysis we estimate that a total of about 5 Mbps (approximately 1.6% of the human genome) is part of at least one such match. The longest matching sequence segments are about 5000 bps in length, suggesting that the majority of segmental duplications probably spawn a few kbps, as reported in previous studies [44, 102].

From Eq. (3.20), one can define the prefactor of the MLD $A = \lambda K/\mu$ so that Eq. (3.20) becomes:

$$M_\infty(r) = \frac{\lambda K}{\mu}\frac{L}{r^3} = \frac{A}{r^3}L. \tag{3.25}$$

From the value of A, one can derive the value of the longest exact match $r_{\max}$ expected in the neutral case,

$$r_{\max} \simeq (AL)^{1/3} \tag{3.26}$$

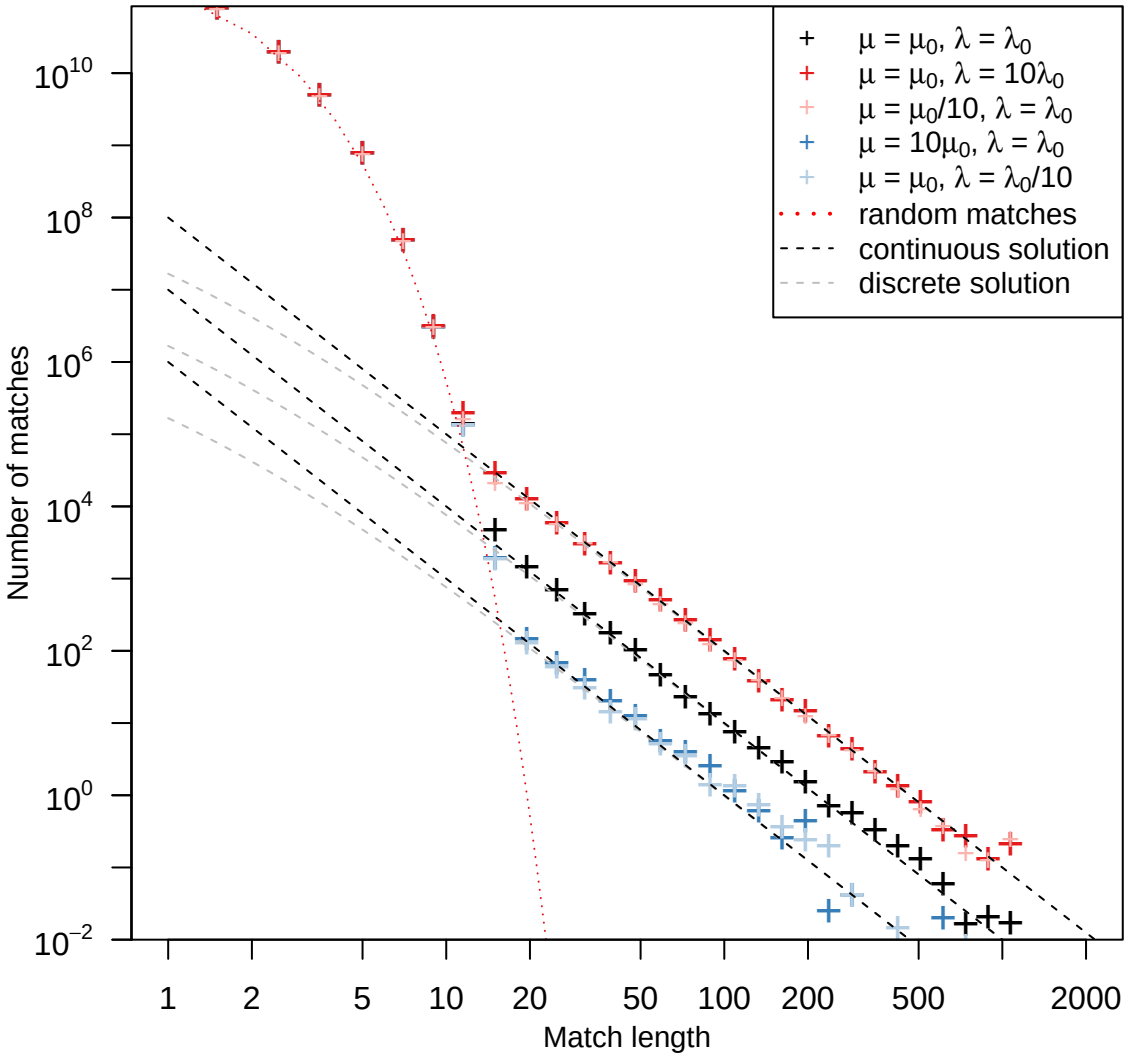

FIGURE 3.3: The match length distribution computed for the self-alignment of sequences simulated using the model described in section 3.2.2 with different values of the duplication rate λ and of the mutation rate μ, with $\lambda_0 = 10^{-3}$ and $\mu_0 = 10^{-1}$. As expected, sequences with the same value of $A = \lambda K/\mu$ exhibit very similar distributions. The red dotted line represent the expected distribution obtained when computing the same experiment on a random iid sequence of the same length and with the same nucleotide frequencies. For small lengths, the MLD is consistent with the random expectation. The dashed line represents the theoretical distribution calculated for 3 different values of the prefactor $A = 100$,10 and 1 in the continuous case (black) or in the discrete case (gray). All MLDs are represented using logarithmic binning.

if $(AL)^{1/3} \leq K$ and $r_{\max} = K$ otherwise. In the human genome, as $A \approx 1$ (see Fig. 3.2), as L, the length of the genome after RepeatMasking, is of the order of 1 Gbp and as the typical size of a segmental duplication, K, is of the order of

10 kbps, the length of the longest expected exact match is $r_{\max} \simeq 1000$ bps. In a random sequence of the same length the value of $r_{\max}$ would only be about 30 bps. Note that $r_{\max}$ is not very sensitive to the values of A and L. For instance, in a genome of the same length but with $A = 0.1$, the value of $r_{\max}$ would just change twofold, resulting in $r_{\max} \simeq 500$.

This prefactor $A = \lambda K/\mu$ can be interpreted as the ratio of the number of nucleotides affected by duplications compared to the number of nucleotides affected by mutations. For the human genome this factor is close to one. This indicates that the amount of information that is "backed up" by segmental duplications is on average equal to the amount that is lost due to mutations. Note, however, that the spatial distribution of segmental duplications in the human genome is very complex and not specific to coding sequences. Therefore, this process might not save coding sequences from deterioration per se. For present day biological evolution, natural selection is probably a more powerful force to maintain and evolve genomic information over long periods of time.

Furthermore, one can also derive an estimate of the value of the duplication rate λ from the value of the prefactor A. Assuming that mutations occur with a rate of about 1.5 per billion years [44, 103] we can easily derive that about 4.5 Mbps of DNA per million years is duplicated in the human lineage. Assuming further that a typical duplication is 10 kbps long, we find that the duplication rate λ is of the order of 1.5×10^{-13} per bp and per year. These estimates agree with the ones given by Bailey and Eichler [44].

Interestingly, when restricting our analysis to the exons of the human genome, we find the same power-law tail with exponent $\alpha = -3$ in the MLD, see Fig. 3.2. This is quite surprising as one would expect that mutations do not occur randomly in exons, due to all the evolutionary constraints that shape the evolution of exons. For instance, the third positions of any exon have a higher mutation rate than the first two positions, due to the redundancy of the genetic code. Some small domains of proteins [104] are also known to be of particular importance, and are, for this reason, under higher evolutionary constraint than the rest of the

protein. Thus, the region of the exon coding for this domain shows a smaller mutation rate than the rest of the exon. For all these reasons, one would expect that the random stick breaking process cannot be applied to the evolution of exons. However, it might be that, over all exons, various local biases compensate for each others. More importantly, here we only study mutations in duplicated exons, which are a subgroup of all exons (representing roughly 50% of the total exome). Especially, mutational constraints are known to change dramatically when a gene is duplicated. Many scenarios are feasible, notably one (referred to as pseudogenization) where constraints on one of the two copies are totally relaxed, and the relaxed copy then evolves according to our model (see for instance Lynch and Conery [42] and see Section 3.4 for a wider discussion on this topic). Note also the different values of the prefactor ($A \simeq 3$ compared to $A \simeq 1$ in the complete genome) in the MLD computed from the self-alignment of exons (see Fig. 3.2). This is most likely due to a lower nucleotide substitution rate in these regions of the human genome.

Finally, we remark that in contrast to three-dimensional objects, which also show scale-invariant behavior in their fragment size distribution when broken [105], our one-dimensional objects, segmental duplications, need to be continuously generated and broken up to give rise to the observed power-law tail as a superposition of exponential distributions for different degrees of fragmentations. This condition of continuity seems to be sufficiently met for segmental duplications in the human lineage. This is not true for repetitive elements, which have been copied into our genome in irregular bursts. Therefore match length distributions of the non-repeat masked genome, which is clearly dominated by inter-repeat matches, does not have a power-law tail with exponent $\alpha = -3$, see Appendix B.

3.2.4 Limitations of the Simple Model

Although this model explains well the behavior observed in several genomes we analyzed, it does not account for all our observations. First, in the genome of several species, the MLD does show a power-law behavior, but with an exponent

$\alpha = -4$, (see for instance the MLD computed from the zebrafish, the rabbit, or the *arabidopsis thaliana* genome on Fig. 3.4). As one characteristic of our model is that the exponent α is always equal to -3, these distributions clearly deviate from our simple model. In some other genomes, the MLD exhibit a totally different behavior (as for instance the genome of the Orangutan, as well as many others), see Fig. 3.15 in Section 3.5.2 or Taillefer and Miller [106].

Second, even for the genomes where our model seems correct, we notice that the assumption that each match has exactly two occurrences is often violated. Indeed, it has been shown that some regions of the human genome (also called duplication hotspots) are much more prone to duplicate than others (duplication coldspots), see for instance Linardopoulou, Williams, Fan, Friedman, Young, and Trask [49], Zhang, Lu, Chung, Yang, and Li [107]. Similarly, the assumption that duplications occur continuously in time might also be an oversimplification. Notably, a burst of segmental duplication events has been reported in the recent history of the hominid lineage, see Marques-Bonet *et al.* [108].

To quantify the importance of these phenomena in eukaryotic genomes, we used self-alignments to perform a different analysis. For each base of the genome, we retrieved the number of maximal exact matches (MEMs) longer than 20 bps in which this base pair was involved (see Fig. 3.5 for an example) referred to as the "coverage" of a base pair in the following, and computed the coverage distribution (CD). We show the CD computed on the RepeatMasked genomes of four species, Human, Rabbit, *Arabidopsis thaliana* and Zebrafish in Fig. 3.6.

If our assumptions of Section 3.2 were fulfilled, we would expect all base pairs to be involved in either 0 or 2 matches. Interestingly, in the genome of the four species we analyzed, we observed that these distributions again exhibit a fat tail. Namely, many base pairs exhibit a very high coverage. This indicates that our previous assumptions leading to $N(\tau) = N_0 \, \forall \tau$ are violated, and that the probability that segments that have already duplicated duplicate again is not negligible.

Moreover, we observe that both the rabbit and zebrafish genomes, have many more bases with a high coverage than the human genome (even though the genome of

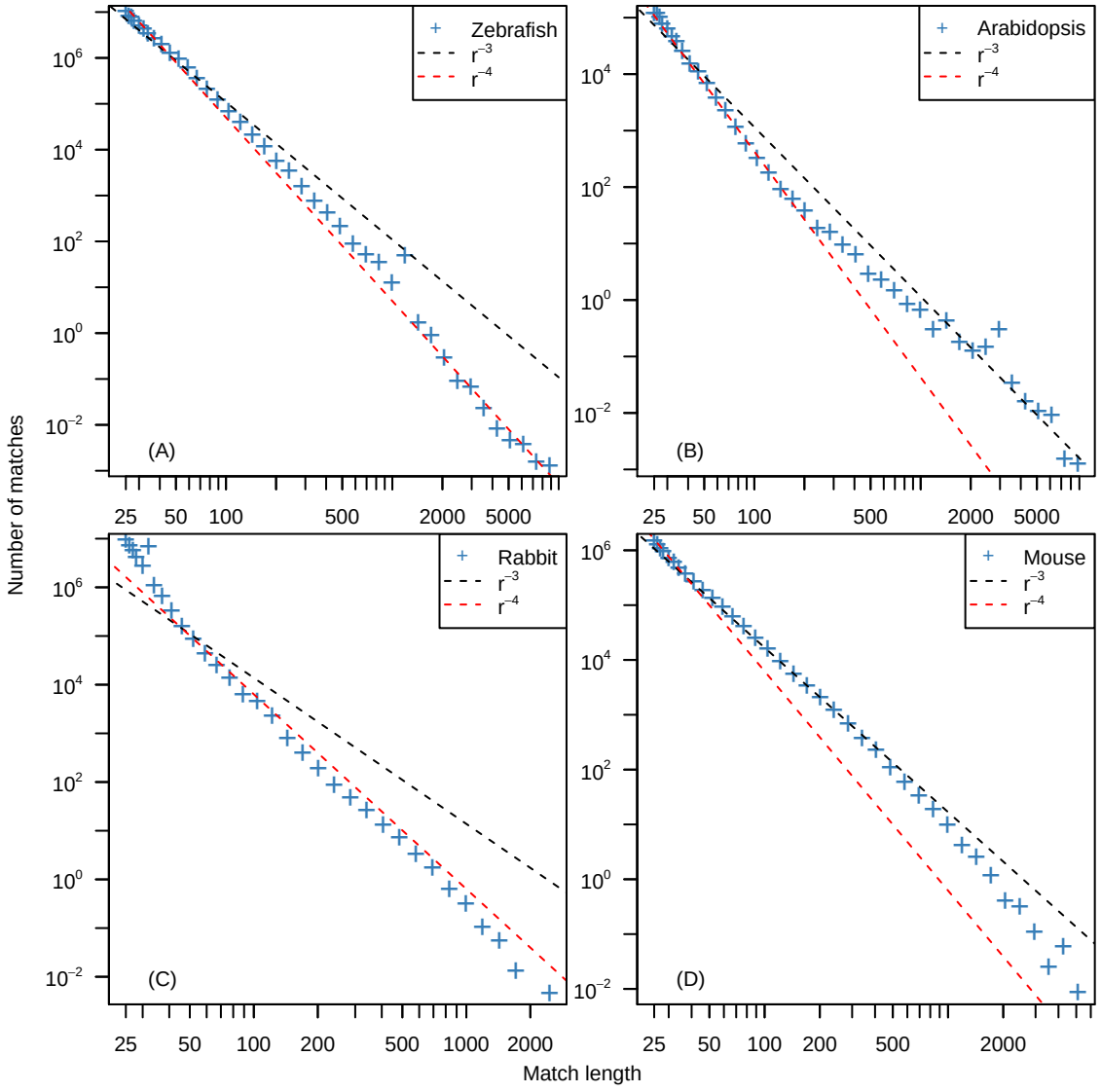

FIGURE 3.4: The match length distribution (MLD) computed for the self-alignment of different species. The dashed lines represent the power-law functions L/r^3 (black) and L/r^4 (red) where r is the match length. All the curves are represented using logarithmic binning. The self-alignment of the Repeat-Masked genome of: (A) the Zebbrafish (*Danio rerio*), (B) *Arabidospis thaliana* (C) the Rabbit (*Oryctolagus cuniculus*) and (D) the Mouse (*Mus musculus*) genome excluding the Y chromosome.

the Zebrafish is much smaller, and the genome of the Rabbit is of the same size than the human genome). This is also true for *Arabidopsis thaliana*, although to a lesser extent. However, arabidopsis genome is much smaller (roughly 10 time smaller) than the human genome, and thus the total number of duplicated base pairs is much lower in Arabidopsis than in Human. Normalizing by the genome length, the proportion of matches with more than two occurrences is much

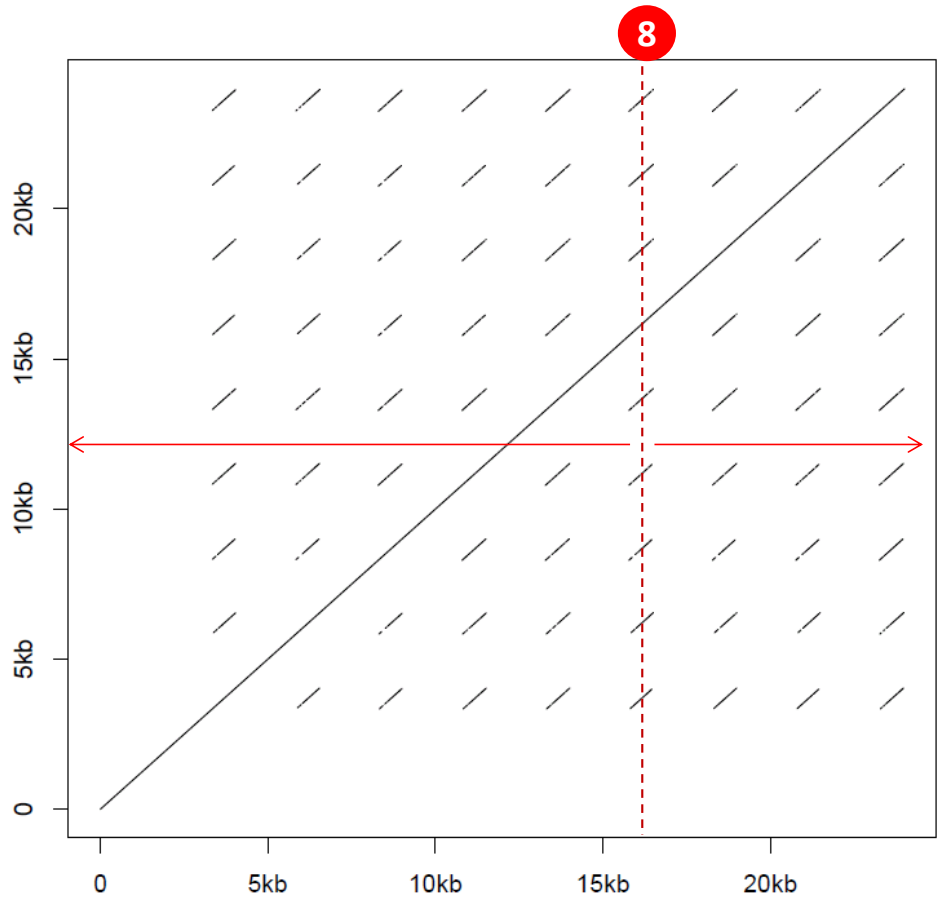

FIGURE 3.5: The self-alignment of 25kbps on the human chromosome X. Black dots in the grid represent exact matches longer than 20bps. Counting the number of time a base pair is involved in a match is equivalent to counting the number of time the red vertical line crosses a black line. To get the full CD, we repeat the same procedure for all base pairs of a genome, and on the entire self-alignment grid.

higher in Arabidopsis than in Human, and the deviation from our hypothesis of Section 3.2 seems stronger for the arabidopsis genome than for the human genome. Thus, this phenomenon might explain why we observe different behaviors for the MLDs of these three species. In the following, we relax the assumptions on $N(\tau)$ and calculate $N(\tau)$ and the resulting $M(r)$ for different biologically relevant evolutionary scenarios.

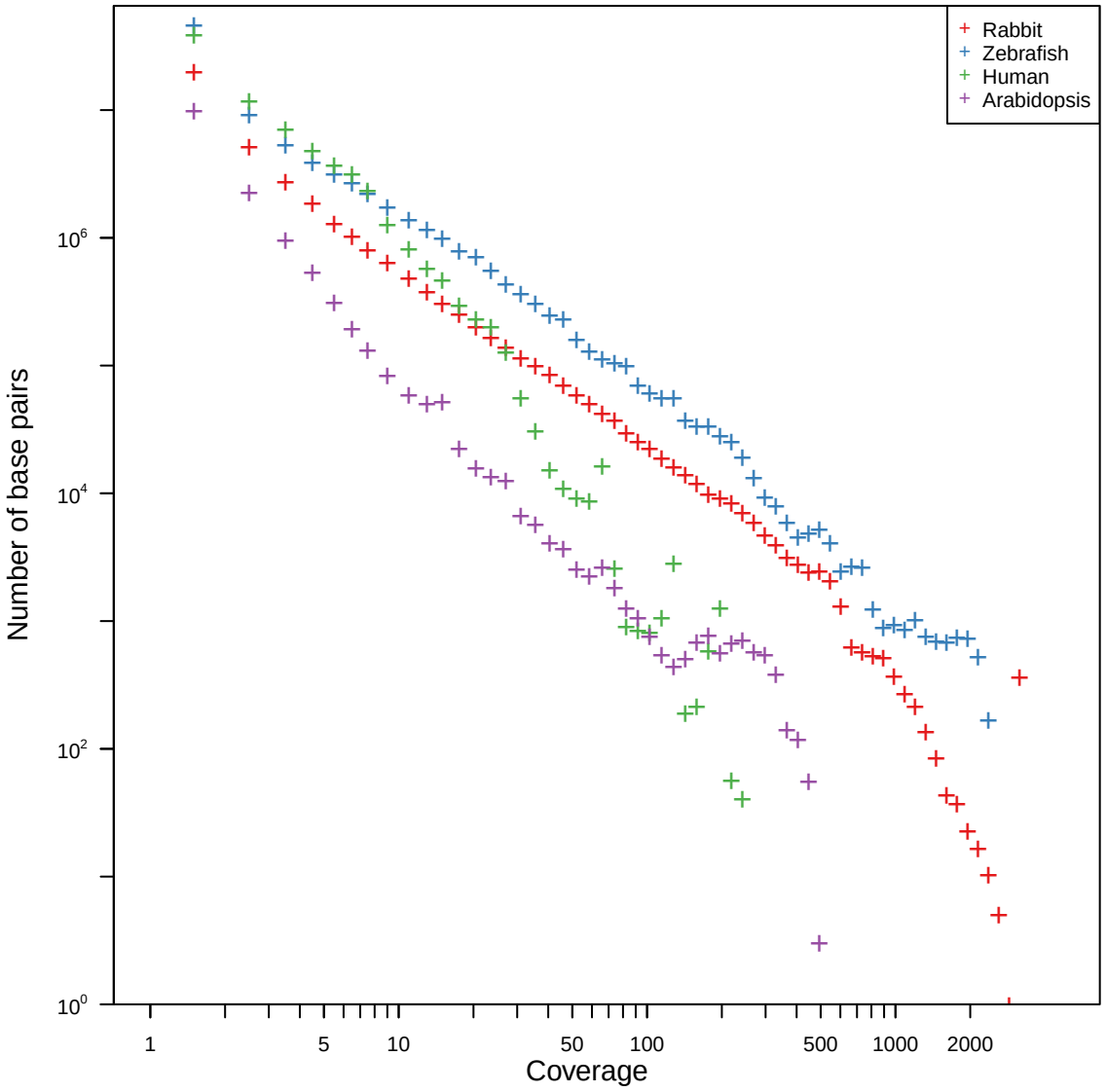

FIGURE 3.6: The coverage distribution computed from the self-alignment of four RepeatMasked genome: Zebrafish, Human, Rabbit and *Arabidopsis thaliana*.

3.3 Yule Trees

3.3.1 Theoretical Calculation

In this section, we study an evolutionary scenario where a particular sequence segment and its duplicated offsprings duplicate again with a fixed duplication rate. With this process, we want to model the case where a particular segment of DNA, or a particular region, has a higher probability to duplicate than the rest of

the genome, and to study whether such a process could lead to the power-laws with an exponent $\alpha = -4$ observed in the MLD of several genomes. This would allow certain regions to be duplicated more than once, and could shape a CD similar to the one observed in real genomes, where several segments are duplicated many times. In this section, we assume that after a duplication, both duplicates can duplicate again, and do so with the same probability. We consider the case where only one of the two duplicates can duplicate in Section 3.4.

According to this process, a segment of length K of the genome, as well as all its offsprings, duplicates again and again with a constant duplication rate per bp λ, such that the duplication rate per segment is λK. The mutation rate per bp μ is the same all over the genome. According to this process, one particular segment at time $t = 0$ gives rise to a family of segments. The evolutionary history of such a family can be well described by a Yule tree (see example in Fig. 2.2), and its size grows exponentially in time.

To calculate the theoretical MLD in this Yule tree scenario, we have to compute the distribution of pairwise distances $N(\tau)$ on such a tree. Let us focus on the case where we start from one ancestral sequence segment, as exemplified in Fig. 2.2.

Let us consider a segment that has evolved according to this process for a time T. Pairs of leaves, separated by an evolutionary time in the interval $[\tau, \tau + d\tau]$, have branched out in the time interval $[T - \frac{\tau+d\tau}{2\mu}, T - \frac{\tau}{2\mu}]$. Now consider the branching process as illustrated in Fig. 2.2. The first branching happened at time T_1 and the two resulting subtrees encompass, say, M_1 and M_2 leaves, respectively. If we define the number of pairs of leaves separated by an evolutionary time τ on a Yule tree given a total time T as $N(\tau|T)$, it follows that:

$$N(\tau|T) = 2N(\tau|T - T_1) + M_1 M_2 \delta(\tau - 2\mu(T - T_1)) I_{(0 \leq \tau \leq 2T\mu)}. \tag{3.27}$$

The first term on the right hand side counts the number of pairs at a distance τ inside each subtree, using the fact that any subtree of a Yule tree is also a Yule tree with the same birth rate. The second term on the right hand side of Eq. (3.27) counts the expected number of pairs at a distance τ from each other between the

two subtrees. The function I is the indicator function defined by:

$$I_{(\text{condition})} = \begin{cases} 1 & \text{if condition holds} \\ 0 & \text{otherwise,} \end{cases} \tag{3.28}$$

and $\delta(x)$ is the Dirac delta function. Averaging over M_1, M_2 (using Eqs. ((2.10) and (2.11)) with time $T - T_1$ and death rate $\delta = 0$) and then T_1, which follows an exponential distribution with mean $1/\lambda$, one obtains:

$$\begin{aligned} N(\tau|T) &= 2\lambda \int_0^\infty N(\tau|T - T_1)\, e^{-\lambda T_1} dT_1 + \frac{\lambda}{2\mu} e^{3\lambda\tau/(2\mu) - \lambda T} I_{(0\le\tau\le 2\mu T)} \\ &= \frac{\lambda K}{2\mu}\, e^{\lambda K T}\, e^{\lambda K \tau/2\mu} \end{aligned} \tag{3.29}$$

for $0 < \tau < 2T$ and $N(\tau|T) = 0$ otherwise. For a detailed and more general derivation of this and other quantities on Yule trees, see Sheinman, Massip, and Arndt [62] in Appendix C .

Substituting Eq. (3.29) in Eq. (3.6) one obtains for the MLD:

$$M(r) = \frac{\lambda K}{2\mu} \exp(\lambda K T) \int_0^\infty \left[2\tau + \tau^2(K - r)\right] \exp\left[\left(\frac{\lambda K}{2\mu} - r\right)\tau\right] I_{(0\le\tau\le 2\mu T)}\, d\tau. \tag{3.30}$$

This integral can also be explicitly calculated, and we find

$$\begin{aligned} M(r) &= \frac{4\exp(K\lambda T) K\lambda\mu}{(K\lambda - 2\mu r)^3} \Big\{ K(\lambda - 2\mu) + \exp(K\lambda T - 2\mu r T) \times \\ &\quad \left[-K(\lambda - 2\mu) + K(\lambda - 2\mu)(K\lambda - 2\mu r)T + \mu(K - r)(K\lambda - 2\mu r)^2 T^2\right] \Big\}. \end{aligned} \tag{3.31}$$

In the limit $rT\mu \gg 1$ and $\lambda K/(2\mu) \ll r < K$, it leads to:

$$M(r) = \frac{\lambda K^2 e^{\lambda K T}}{\mu} \frac{1}{r^3}. \tag{3.32}$$

Note that the condition $rT\mu \gg 1$ is quite natural as we are interested in long matches (i.e. $r \gg 1$) and in the behavior of the stationary distribution (i.e. after a long time). The second condition implies that the value of the prefactor $A = \lambda K/\mu$ stays small, which is what we observe in real genomes.

In this scenario, the size of the genome grows exponentially in time, and $L = K\exp(\lambda KT)$. Replacing the value of L in Eq. (3.32) we find:

$$M(r) = \frac{\lambda KL}{\mu r^3}. \tag{3.33}$$

Surprisingly, this is exactly the result found in the previous section (see Eq. (3.20)).

Above we focused on the case where, at the beginning of the process, only one segment of the genome can duplicate. We can easily extend our calculation to the case where any segment of the genome can be duplicated. For a single Yule tree, the total length of the segments after time T is given by $Ke^{\lambda KT}$ on average. We now assume that the genome is composed of n duplicating non-homologous segments, and that each segment i evolves independently according to the process described above. If we then denote the duplication rate of the sequence i by λ_i, the total number of matches in the self-alignment of one genome is given by:

$$M(r) = \sum_{i=1}^{n} \frac{\lambda_i K^2 e^{\lambda_i KT}}{\mu} \frac{1}{r^3} = \frac{K}{\mu r^3} \sum_{i=1}^{n} \lambda_i K e^{\lambda_i KT}. \tag{3.34}$$

The total length of the genome at time T is given by:

$$L = \sum_{i=1}^{n} K e^{\lambda_i KT}. \tag{3.35}$$

If all segments have the same duplication rates $\lambda_i = \lambda$, we can substitute Eq. (3.35) in Eq. (3.34), and obtain:

$$M(r) = \frac{\lambda KL}{\mu} \frac{1}{r^3} = A\frac{L}{r^3}, \tag{3.36}$$

with $A = \lambda K/\mu$. And again, we retrieve exactly the solution obtained in the previous section, see Eq. (3.20).

3.3.2 Simulations

To confirm our theoretical results, we simulated sequences evolving according to the process described above. In this version, we start with one random iid sequence of length K. Compared to the previous model, the mutation process stays unchanged, and we denote by μ the mutation rate per bp. We start our procedure by duplicating this first segment to the site adjacent to the right, such that the starting position of the duplicated segment is $c = 0$ and it gets copied to the position $v = K$. The total number of segments, n, of the sequence is then $n = 2$. For subsequent duplication events, we choose one of the n pre-exiting segments, copy it at the end of the sequence, i.e. at position $v = nK$ and increment n by one afterwards. The duplication rate per gene is λK so that the duplication rate per bp is, like in the first model, equal to λ. Note that in this version, the size of the sequence L grows exponentially with time, so we cannot reach a stationary state, and we stop the procedure after a fixed evolutionary time T. Another difference with the preceding model is that duplications can only occur at some fixed positions 0, K, $2K$, ..., nK corresponding to the beginning of a gene. This does not change the expected behavior of the MLD, but allows us to reconstruct easily the relationships between the different genes. The last position of each gene is set to always be an "N" and cannot mutate, so that we do not create chimeric matches spanning the end of one gene and the beginning of its neighbor.

We show the pairwise distance matrix and the corresponding tree for one family generated using this process on Fig. 3.7. The MLD of sequences simulated using this process are in good agreement with our theoretical calculations, see Fig. 3.8.

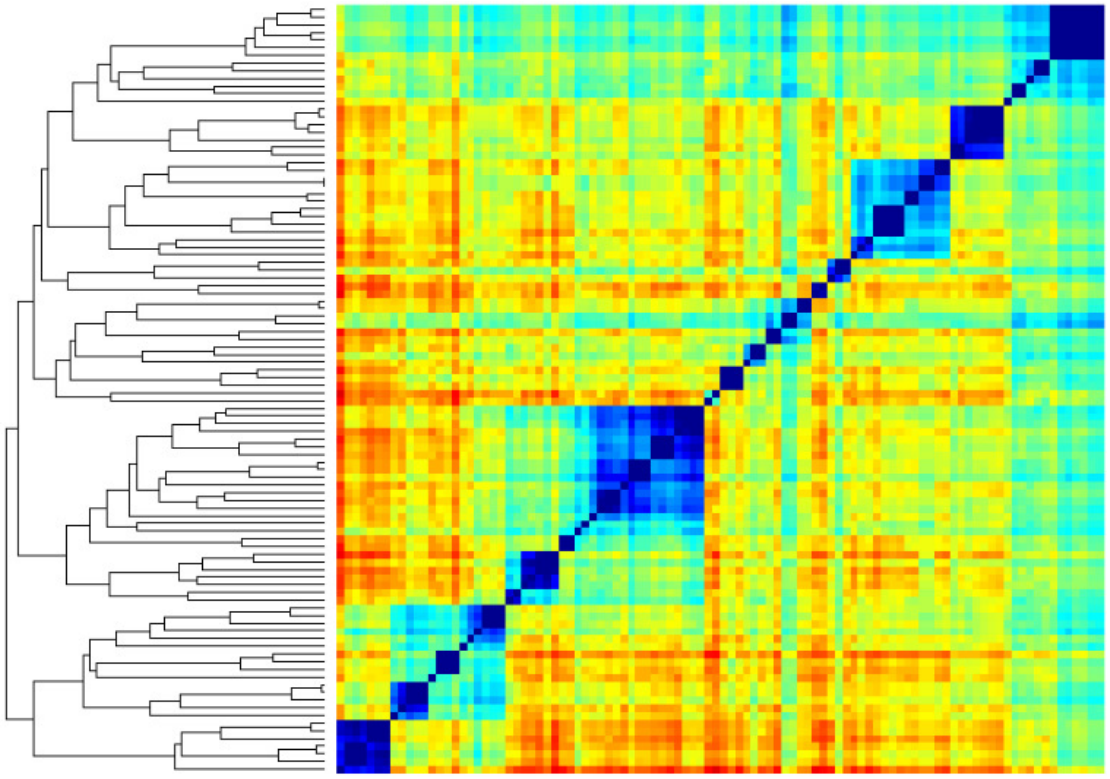

FIGURE 3.7: Tree and distance matrix of a family of simulated sequences for the case where all sequences duplicate with the same duplication rate, λ, and mutate with the same rate, μ, giving rise to a Yule tree. Small distances are depicted in blue, and large distances are in red.

3.3.3 Discussion

Surprisingly, using the MLD alone, one cannot distinguish between the two scenarios, in which either all sequence segments duplicate randomly or only a subset of sequences duplicate presumably many times. Moreover, if the duplication rates of the different duplicating segments of the Yule tree scenario (i.e. the duplication rates in different gene families) are similar, even the prefactor of the MLD are equal in both scenarios. However, the two models differ on the number of occurrences they predict for each match. The model developed in section 3.2 predicts that all matches have 2 occurrences, or at least that the number of matches with more than 2 occurrences is extremely low, while most of the matches of the Yule model have a larger number of occurrences. This second case is probably closer to reality, as argued in Sindi, Hunt, and Yorke [68] and Kim *et al.* [47]. Such a model could also be consistent with the result of our coverage experiment

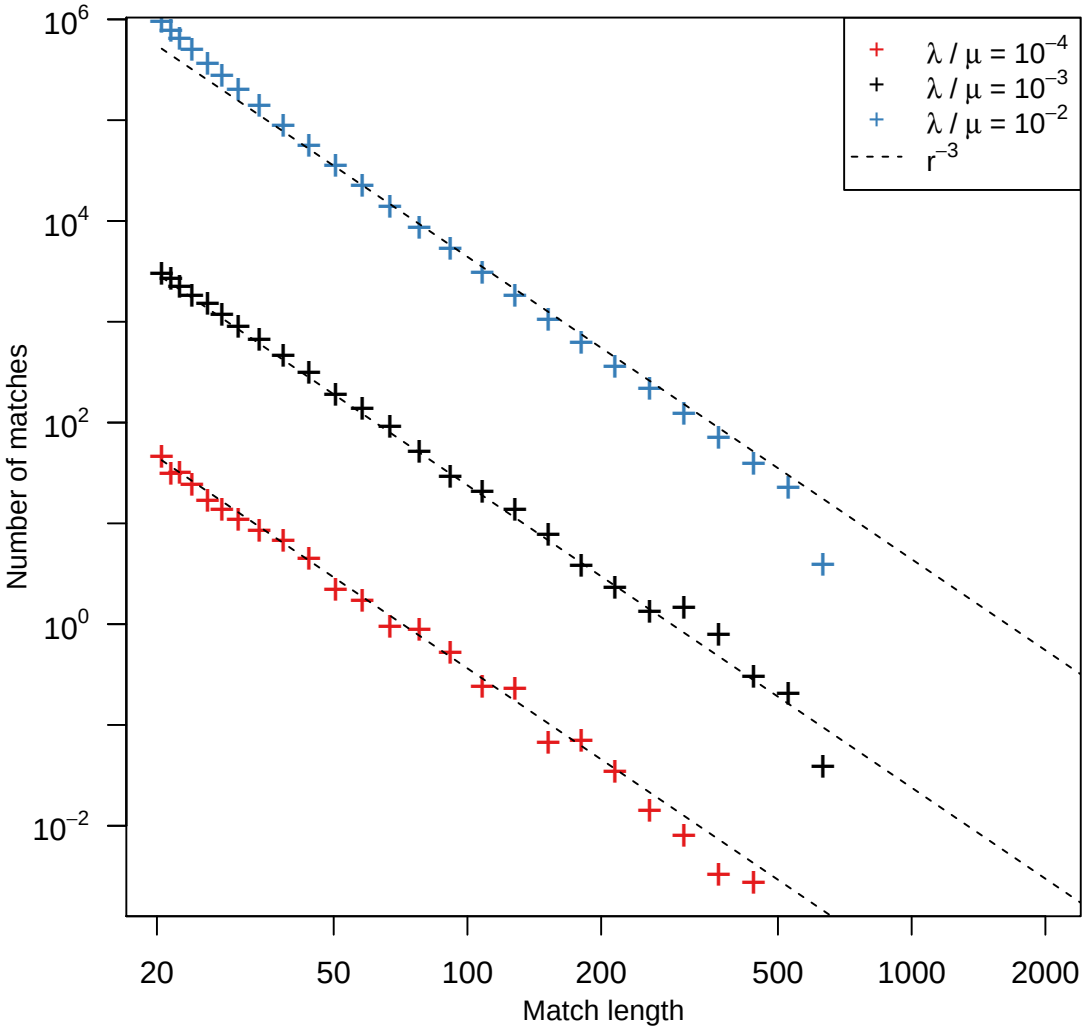

FIGURE 3.8: The MLD computed for the self-alignment of sequences simulated using the model described in section 3.3.2 with different values of the ratio λ/μ, 10^{-2},10^{-3} and 10^{-4}. All three curves result from the simulation of 100 families with $K = 1000$, for a time $T = 5$. All MLDs are represented using logarithmic binning.

shown on Fig. 3.6. However, CD obtained from the self-alignment of sequences simulated using the Yule model with parameters that fulfill the conditions where the power-law is observed (i.e. $\lambda K/2\mu \ll r$) exhibit few bases with a really high coverage. Moreover, the CD decreases exponentially, and thus much faster than what we observe in the CDs of real genomes. Increasing the ratio $\lambda K/2\mu$ leads to a CD consistent with empirical observations, but to a MLD which does not exhibit a power-law distribution anymore. Thus, a more detailed analyses of these CD could help understanding the $\alpha = -4$ power-laws observed.

Note that in this scenario, the branching process we considered is a pure birth process. We can further consider a process where DNA segments are also deleted with a rate δ per bp. In that case, our model becomes a birth death process. Using arguments similar to the one used in the pure birth case, we can calculate the value of $N(\tau|T)$ for this process. Interestingly, as long as the death rate stays smaller than the birth rate, the value of $N(\tau|T)$ in the birth death process is equivalent to the value found for a pure birth process with a birth rate equal to $\lambda - \delta$, such that:

$$N(\tau|T) = K\frac{\lambda - d}{2\mu}\exp((\lambda - \delta)\,KT)\exp\left(\frac{(\lambda - \delta)K\tau}{2\mu}\right), \tag{3.37}$$

see Appendix C for the detail of the calculation. Thus, only the prefactor of the MLD would change, and the exponent of the power-law distribution would still be $\alpha = -3$.

In this section, we have shown that allowing duplicates to duplicate again does not change the exponent α of the power-law obtained in the MLD of the self-alignment of a genome. Thus, we still have no explanation for the observation of MLDs with a different exponent obtained from several genomes (see Fig. 3.4). In the next section, we study another mechanism of gene duplication. We show that this mechanism leads to a different topology of the gene family tree, and that it can explain the observation of MLDs exhibiting power-laws with an exponent $\alpha \neq -3$.

3.4 The Case of Retroduplication

3.4.1 Theoretical Calculations

Segmental duplication is not the only biological process that produces duplications in eukaryotic genomes. Retroduplication is a well known biological mechanism which consists in the reverse-transcription of a mature mRNA molecule (i.e. after splicing of its introns), into the genome. For this reason, it generates partial duplicates. As retroduplicants also do not contain regulatory elements and promoters,

they mostly produce non-functional copies, highly similar to the gene transcript, commonly known as processed pseudogenes [52, 53]. Various functions have been found for several such processed pseudogenes, and the debate about the potential role of these duplicates is still open, see for instance Kaessmann, Vinckenbosch, and Long [53] or Okamura and Nakai [54], but it seems that most of the time, they result in non-functional evolutionary dead-ends.

To study the relationship between the sequences resulting from such process, we focused on the large family of 113 processed pseudogenes stemming from the retroduplication of the gene coding for the ribosomal protein RPL21 in the human genome. We present the resulting pairwise distance matrix and a compatible phylogenetic tree in Fig. 3.9 (see details in Section 2.5). In contrast to the previous scenario which generates Yule trees (Section 3.3), our results on RPL21 suggest that all these pseudogenes were actually generated by reverse-transcription of a single functional transcript.

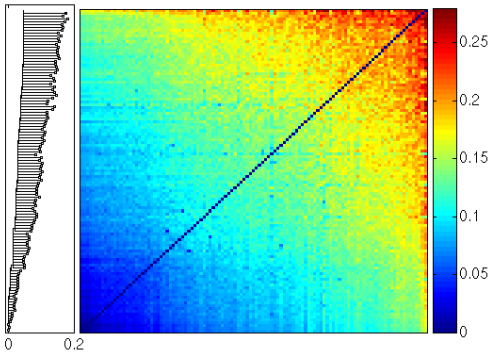

FIGURE 3.9: Distance matrix representing all pairwise distances computed from the 113 processed pseudogenes of the RPL21 gene and the corresponding phylogenetic tree. The rows and the columns of the distance matrix are sorted with respect to their average. The resulting order is used to constrain the topology of the phylogenetic tree (see details in Section 2.5). Small distances are depicted in blue and large distances in red.

According to this mechanism, a gene of length K duplicates with rate λK, while its duplicates (processed, non-transcribed pseudogenes) do not duplicate. Since the selective pressure is expected to be much weaker on pseudogenes (if any) than on the gene, we assume that the gene exhibits a much lower effective mutation rate than its pseudogenes. This results in a tree similar to the one shown in Fig. 3.10.

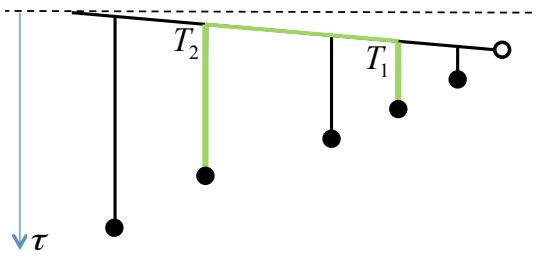

FIGURE 3.10: An example of the rooted tree of a pseudogene family (filled circles) stemming from one gene (open circle). The gene evolves much slower than its pseudogenes, and the pseudogenes do not duplicate. The evolutionary distance between two leaves (green path) is the sum of the evolutionary distance covered by each pseudogene since its retroduplication event and the evolutionary distance covered by the gene between the two retroduplication events. All circles represent contemporary sequence segments.

The evolutionary time that separates two leaves on such a tree is a sum of three evolutionary times: the evolutionary time elapsed after the first retroduplication event, the evolutionary time elapsed after the second retroduplication event and the evolutionary time elapsed in the source gene between the two retroduplications (see the green path on the tree of Fig. 3.10). Defining μ as the mutation rate of a pseudogene and μ_S as the mutation rate of the source gene, the evolutionary time separating two randomly chosen retroduplicants is given by:

$$\tau = \mu\left(T - T_1\right) + \mu\left(T - T_2\right) + \mu_S\left|T_1 - T_2\right|, \tag{3.38}$$

where T_1 and T_2 are the times of the two retroduplications. Assuming a uniform distribution of T_1 and T_2 between 0 and T, the density of pairs of pseudogenes separated by an evolutionary time τ after time T is given by averaging Eq. (3.38) over T_1 and T_2:

$$N\left(\tau\right) = \int_0^T \int_0^T \frac{dT_1}{T}\frac{dT_2}{T}\delta\left(\tau - \left[\mu\left(T - T_1\right) + \mu\left(T - T_2\right) + \mu_S\left|T_1 - T_2\right|\right]\right), \tag{3.39}$$

where δ denotes the Dirac function. This integral is easy to calculate and results in :

$$N(\tau) = \begin{cases} \dfrac{\lambda^2 K^2}{2\mu^2} \dfrac{1}{1+a}\,\tau & \text{for } 0 \leq \tau \leq (1+a)\mu T \\ \dfrac{\lambda^2 K^2}{2\mu^2} \dfrac{-1}{1-a}(\tau - 2\mu T) & \text{for } (1+a)\mu T \leq \tau \leq 2\mu T, \end{cases} \tag{3.40}$$

where $a = \mu_S/\mu$ and is assumed to be smaller than one.

This is a continuous piece-wise linear function, which vanishes for $\tau = 0$, namely $N(0) = 0$. It increases linearly with τ for small values of τ, reaches a maximum at $\tau = (1+a)\mu T$ and then decreases linearly with τ, vanishing for $\tau \geq 2\mu T$. Such a qualitative trend can be observed in the data for RPL21 pseudogenes shown on Fig. 3.9: the number of entries showing a small distance in the distance matrix is small, it increases with the distance, reaches a maximum around 0.12 and then decreases for higher distances.

Substituting Eq. (3.40) in Eq. (3.6), one obtains in the limit of $rT\mu \gg 1$ and $0 < r \ll K$ the following distribution for the tail of the MLD:

$$M(r) = \frac{3K^3\lambda^2}{(1+a)\,\mu^2} \frac{1}{r^4}, \tag{3.41}$$

i.e. a power-law with exponent $\alpha = -4$.

This result suggests that the self-alignment of processed pseudogenes (retroduplicants) is expected to generate an MLD distributed as a power-law with exponent $\alpha = -4$. To confirm this prediction, we concatenated all the annotated processed pseudogenes of the human genome, to construct the so-called human "processed pseudogenome".

First, we downloaded the sequence of all 16889 known pseudogenes of the human genome from the `pseudogene` database [109]. We then filtered these sequences according to their annotation in this database, keeping only those annotated as processed pseudogenes (9053 pseudogenes left). Using the positions of these different pseudogenes in the genome, we ensured that the different pseudogenes were not overlapping in the human genome. When this was the case (only 25 times), we concatenated the two sequences into one longer sequence overlapping the two

pseudogenes. We then concatenated all the remaining sequences into one long sequence of 6433 kbps. To separate the different pseudogenes, we added a letter 'N' between each pseudogene, to avoid creating artificial and irrelevant matches.

Finally, we computed the MLD from the self-alignment of this processed pseudogenome. It shows a good agreement with the prediction of Eq. (3.41), see Fig. 3.11, although we can observe a significant deviation from the prediction at the very end of this MLD. This deviation could be explained either by subsequent segmental duplications of retroduplicated loci or by selective constraints on the retroduplicates making them more conserved than expected by our neutral model.

3.4.2 Simulations

To confirm our theoretical results, we simulated sequences evolving according to the process described above. In this version, we first create one random iid gene sequence S_g of length K. Point mutations in this gene occur with a mutation rate μ_S.

We start our procedure by duplicating this first gene to create a new pseudogene sequence, S_p. Point mutations in the pseudogene sequence occur with a mutation rate μ. The gene sequence is the only one able to duplicate. When a duplication occurs, a copy of the gene is added after the last position of the pseudogene sequence, i.e. $v = nK$, and we increase n, the number of pseudogenes of the sequence, by one. As in the Yule tree version, the length of the pseudogene sequence L increases with time, but in the present case, the growth is linear in time. We stop the procedure after a fixed evolutionary time T. Again, the last position of the gene is set to always be an "N" and cannot mutate, so that we do not create matches spanning the end of one gene and the beginning of its neighbor. At the end of the simulation, we only retrieve the pseudogene sequence S_p and discard the gene.

We show the distance matrix representing the pairwise distances between all pairs of pseudogenes and the resulting tree for one simulated sequence on Fig. 3.12 . As

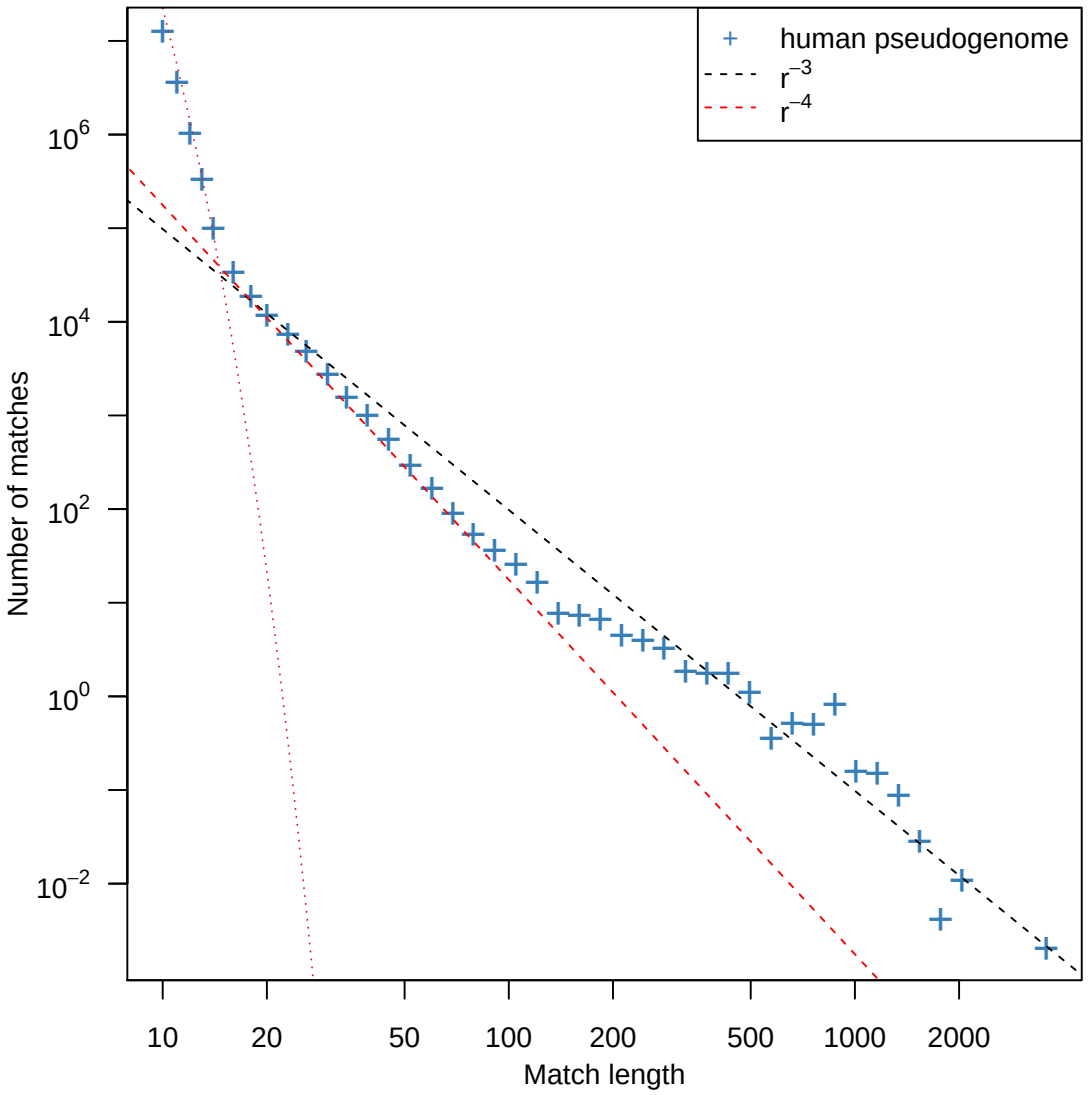

FIGURE 3.11: The MLD computed from the self-alignment of the human processed pseudogenome. The total length of this genome is $L = 6,433,368$ bps. The red dotted line represents the expected distribution for a random sequence of the same length, and the red and black dashed lines represent power-laws with exponent $\alpha = -4$ and $\alpha = -3$ respectively.

expected, the tree and the distance matrix are in good agreement with the empirical experiment conducted on the RPL21 pseudogene family shown on Fig. 3.9. We show the MLD resulting from 100 iterations, and for different parameters on Fig. 3.13. The simulated MLDs show a good agreement with our theoretical calculations, and the MLD computed from the self-alignment of the human processed pseudogenome.

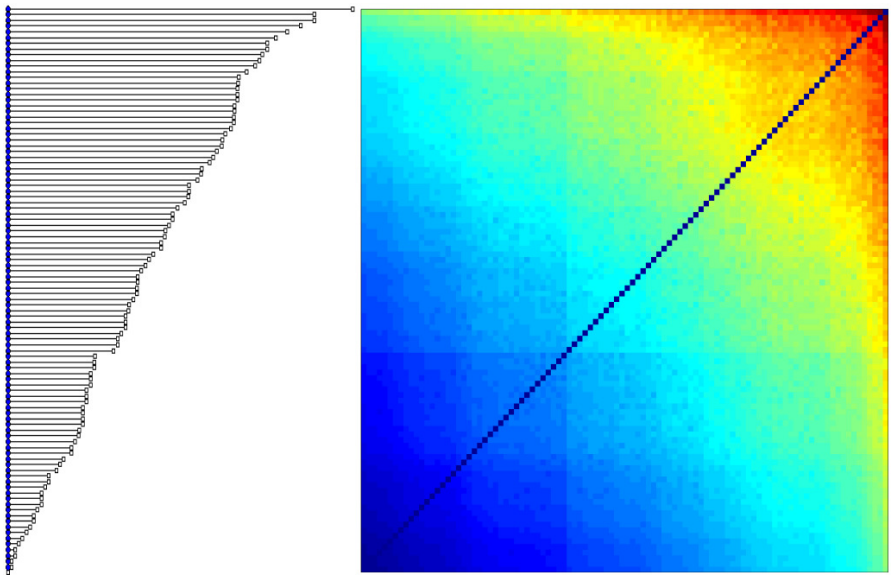

FIGURE 3.12: Tree and distance matrix of a family of simulated sequences for a simple case using model of section 3.4.2 and with the mutation rate in the gene μ is set to 0. Small distances are depicted in blue, and large distances are in red.

3.4.3 Discussion

In this section, we have shown that a sequence duplicating through retroduplication exhibits a particular gene tree topology. Calculating the value of $N(\tau)$ on such a tree, we have shown that the MLD of genomes where this process is active exhibit an $\alpha = -4$ power-law distribution.

In real genomes, both processes (segmental duplication and retroduplication) are active, and the behavior of the MLD observed in any genome depends only on the dominating duplication process. If most of the duplicates of a genome are retroduplicates (i.e. if the retroduplication rate is higher than the segmental duplication rate), we expect the MLD of the self-alignment of this genome to exhibit an $\alpha = -4$ power-law, while if they are segmental duplicates, we expect an $\alpha = -3$ power-law. Note that for equal rates, as $r^{-4} \ll r^{-3}$, the segmental duplication is supposed to dominate, especially for long lengths. Hence, for the retroduplication process to dominate, it requires the retroduplication rate to be much higher than the SD rate.

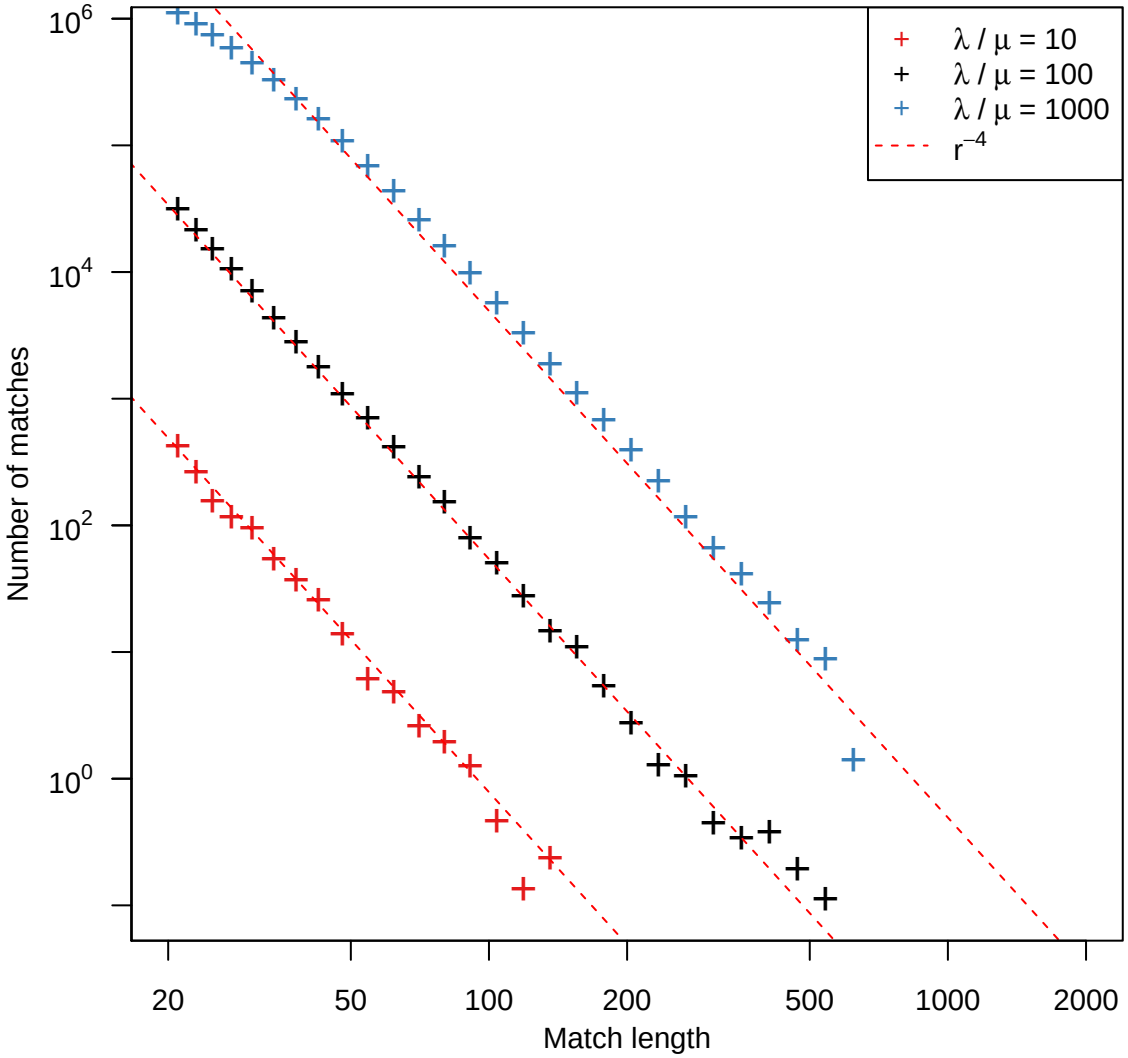

FIGURE 3.13: The MLD computed for the self-alignment of sequences simulated using the model described in section 3.4.2 with different values of the duplication rate λ and of the mutation rate μ . For simplicity, the mutation rate in the gene is set to be $\mu_s = 0$ in these simulations. All MLDs are represented using logarithmic binning. Each MLD was obtained by averaging over 100 simulations with $\lambda = 1$ per unit of time, $T = 100$ and $K = 1000$.

However many duplication scenarios could lead to an $\alpha = -4$ power-law, as we will show in the next chapter. For this reason, the appearance of an $\alpha = -4$ power-law in the self-alignment of a genome does not automatically imply a high retroduplication rate in this genome.

Among the possible scenarios, two of particular biological relevance are the silencing of the segmental duplication process in recent evolution of a genome, or a

recent whole genome duplication (Chapter 5).

3.5 Biological Insights from our Models

3.5.1 Using the MLD to Infer Information on Different Duplication Mechanisms

We have shown that the appearance of a power-law in the distribution of the number of exact maximal matches obtained from the self-alignment of a genome could be explained by the interplay of two basic and neutral evolutionary mechanisms, point mutations and duplications. We have also shown that the exponent of the distribution depends on the properties of the dominant duplication mechanism at work.

Interestingly, the MLD obtained from the human (or mouse) self-alignment (see Figs. 3.2 and 3.4) agrees well with an $\alpha = -3$ power-law distribution, indicating that over all processes generating self-similarities in the human (and Mouse) genome, the dominant mechanism is the segmental duplication of random sequences of the genome. This observation also implies that this process occurred continuously and with a constant rate in the history of these species, and is an ongoing process. As the exponent of the MLD only depends of the dominating duplication process, the observation of an $\alpha = -3$ power-law does not mean that other duplication processes – such as retroduplication – are not at work as well in these genomes. Indeed, we were able to isolate a subset of the human genome where the retroduplication process was dominant, and whose MLD exhibits an $\alpha = -4$ power-law.

In contrast to Human and Mouse, the MLD computed from the self-alignment of the rabbit genome exhibits an $\alpha = -4$ power-law. This could be due to a higher rate of retroduplication in this particular genome. However and as stated previously, many duplication scenarios could lead to an $\alpha = -4$ power-law, and

further analysis are requirred to decide which one is responsible for this behavior. One possible scenario could be the silencing of the segmental duplication process in recent evolution of a genome.

A power-law with exponent $\alpha = -4$ has also been observed in the MLD computed from the self-alignment of the Zebrafish and of the plant model organism, *Arabidopsis thaliana*. However, it has been shown already that a whole genome duplication event occurred recently in those genomes [110, 111]. Such events could have important consequences on the MLD, as we will show in the Chapter 5.

MLDs computed from the self-alignment of many genomes have been presented by Taillefer and Miller [106]. These MLDs exhibit power-laws with various exponent (from $\alpha = -2$ to $\alpha = -4.5$), and some are exponentially distributed. However, genomes with long and highly similar sequences, which are generated by segmental duplications, and especially tandem duplications, are not easy to sequence and assemble when using short read obtained from next generation sequencing technologies (where the typical size of a read is of the order of one hundred bps). As the power-law behavior only holds for long matches — typically longer than the read length — such power-law behavior often remains highly questionable unless the quality of the genomic assembly is high, i.e. comparable to the one of the human and Mouse genomes. When computing a MLD for a new genome, one would expect to obtain a distribution close to an $\alpha = -3$ power-law. Any deviation from this behavior could in principle be interpreted as a lack of proper repeat-masking (notably if one observes peaks for certain lengths in the MLD), a prevalence of another biological process (if one observes a power-law with a different exponent) or a poor assembly quality (if one observes a strong deviation from the power-law behavior). Computing the MLD of a genome, which is a simple and fast computational procedure, can in this sense be of great help in order to understand the biological processes that shape the evolution of this genome, and to assess the quality of its assembly and of its RepeatMasking, as we show with the example of two primate genomes (Orangutan and Macaque) in the following subsections.

3.5.2 Assessing the Quality of the Assembly: Orangutan Example

The MLD obtained from the self-alignment of the Orangutan genome (see Fig. 3.15) has a clearly different behavior than the other species we have studied: the distribution exhibits an exponential decay. According to the broken stick model, an exponential distribution is expected to appear in the distribution computed from one stick which underwent many breaks, (or from a mixture of many sticks which underwent the same number of breaks). In a genomic context, such a distribution could be indicative of a burst of duplications at one particular time point in the evolutionary history of a genome, if the duplications that occurred during this particular period dominates the distribution we observe.

However, results on the MLD of the Orangutan genome should be viewed with caution, as the quality of the assembly of primate's recently sequenced genomes might be quite poor. More specifically, we believe that a few small regions, where the duplication rate is really high, are responsible for the signal we observe. These regions, as well as tandem duplications (duplications where the two matching segments are next to each other in the sequence) are really difficult to assemble [112]. To get rid of biases due to missassembled tandem duplications, we filtered out from our analyses tandem duplications. To do so, we computed the physical distance D, defined as the number of nucleotides separating the two matching segments, for all matches of the orangutan genome which are on the same chromosome. For matches on different chromosome, we set $D = \infty$. We then removed all the matches whose value of D was below a fixed threshold $D_{\max}$. This is equivalent to removing all matches which are close to the principal diagonal on a dot plot, as presented on Fig. 3.14. For this reason, we refer to matches close to one another as matches of the "extended diagonal" in the following.

We then computed MLDs for different values of $D_{\max}$. Interestingly, when we removed the extended diagonal of the orangutan self-alignment, for $D_{\max} > 5$kbps, we retrieved the expected $\alpha = -3$ power-law, see fig 3.15. On the other hand, the

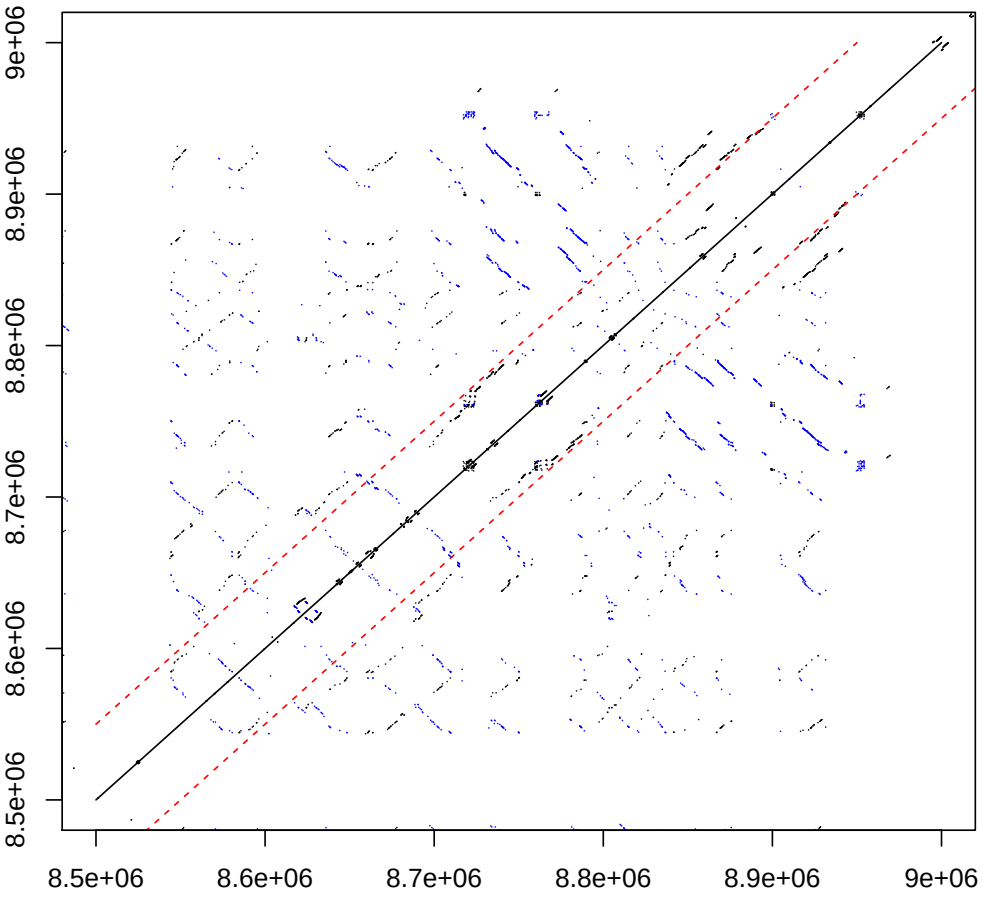

FIGURE 3.14: The Dot plot of a small region (500kbps) on chromosome 19 of the Orangutan RepeatMasked genome. Black dots represent matches with the forward strand, and blue dots matches with the reverse strand. The region we excluded from our analysis lies between the two red dashed lines. Here, the distance $D_{\max}$ separating the two red lines is equal to 10kbps.

MLD computed for matches belonging to the extended diagonal only exhibits a clear exponential distribution (for $D_{\max}$ up to 10 kbps).

As the Orangutan is the only species whose genome exhibit such a shift, one can wonder whether this signature is due to a specificity in the mechanisms generating duplications in this species, or whether assembly artifacts generate these long well conserved matches. Deciding between these hypotheses from the analyses of the sequence only is not easy (especially as this genome was sequenced quite recently

[113]). Note that in some other genomes (as for instance the chicken genome or the coelacanth genome (*Latimeria chalumnae*)) where the MLDs did not exhibit a power-law distribution (most likely because the sequence quality is too low in these genomes), using the same procedure did not change the shape of the MLD.

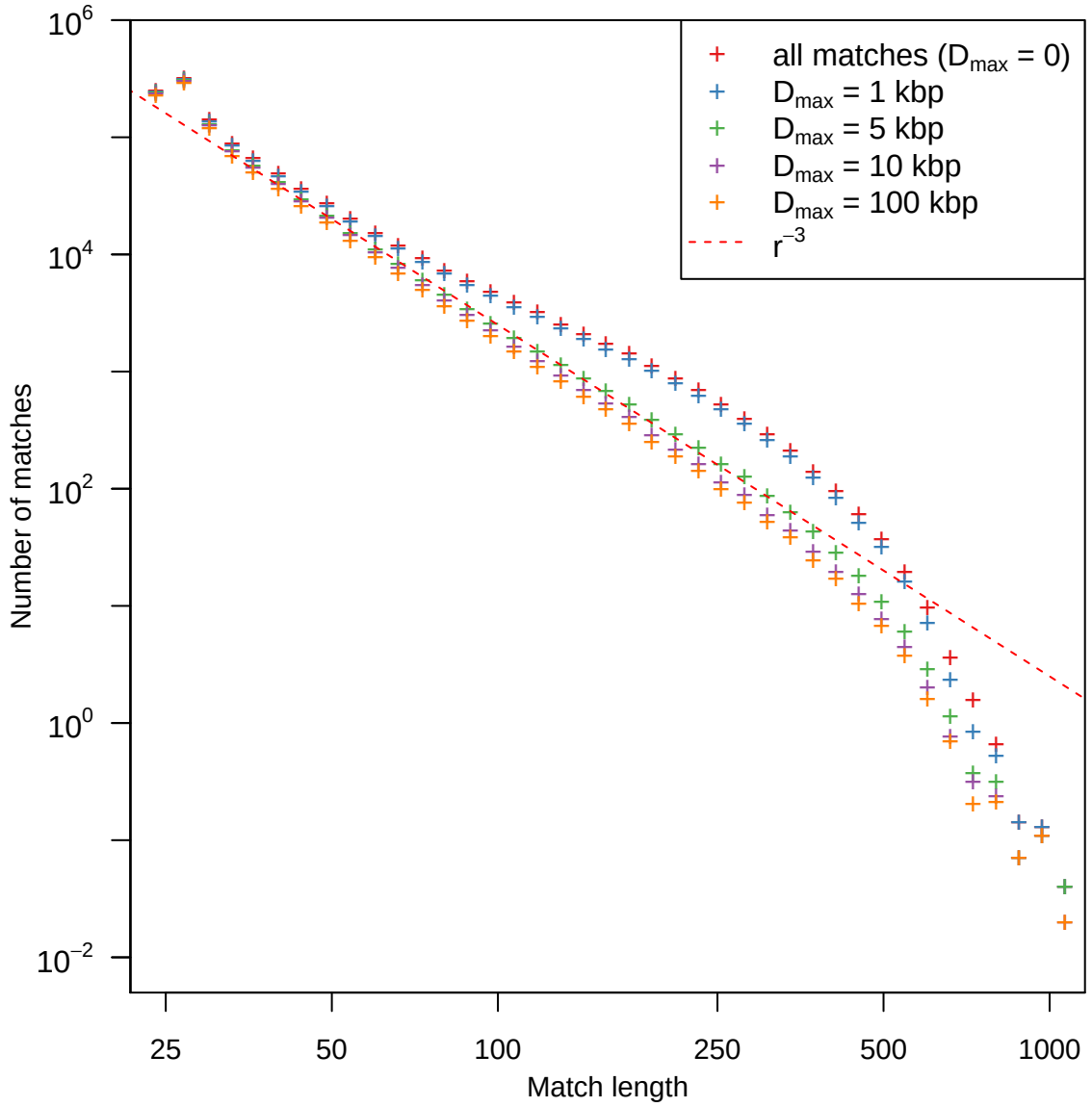

FIGURE 3.15: MLD from the self-alignment of the Orangutan genome where the matches belonging to the extended diagonal have been removed for different values of D. The exponential behavior of the MLD vanishes when we remove matching segment for which $D < 5$kbps.

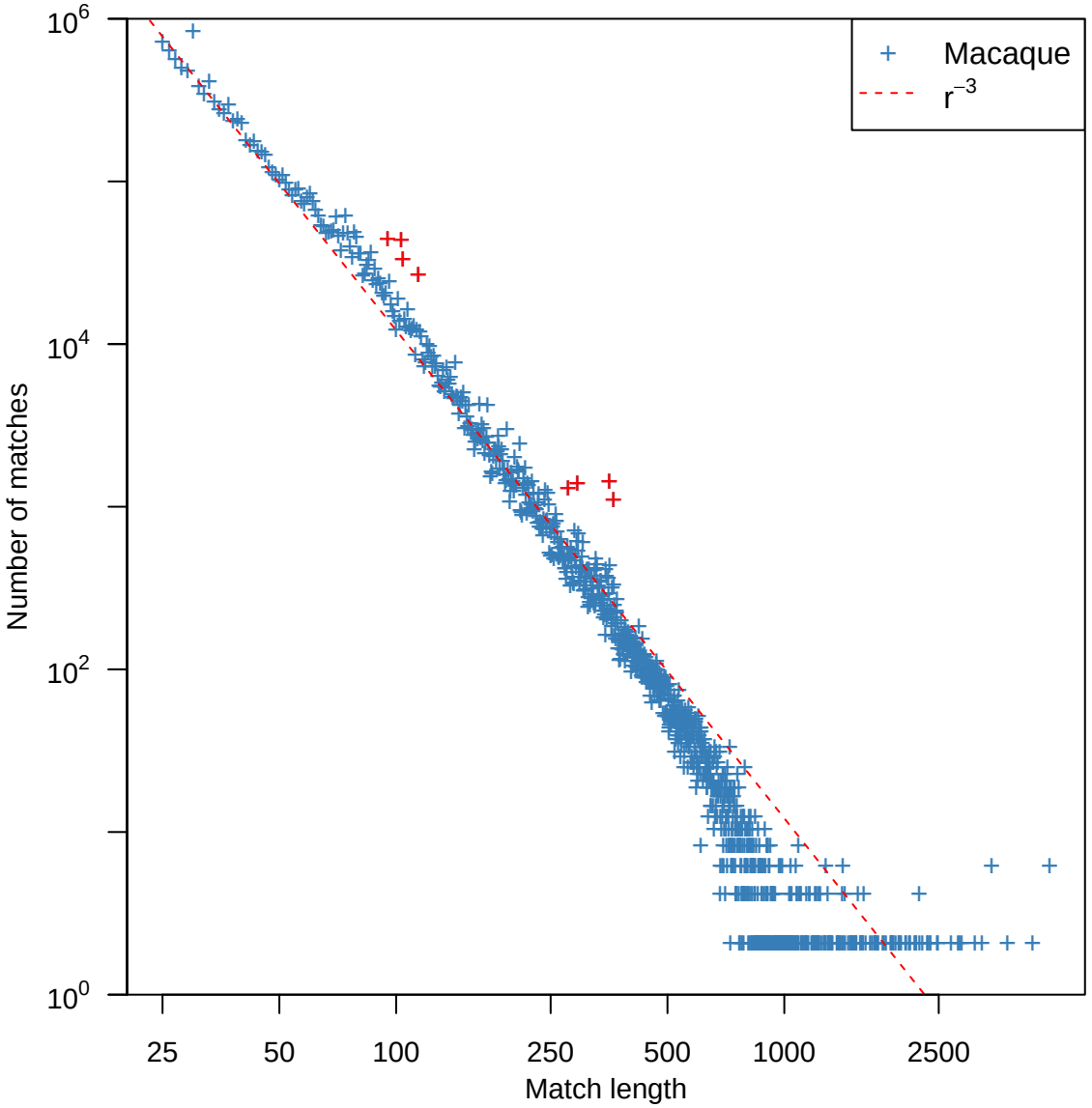

FIGURE 3.16: MLD computed from the self-alignment of the Macacque genome. Here we present the distribution without logarithmic binning to underline the high values obtained for eight specific lengths. These eight outliers are depicted in red.

3.5.3 Assessing the Quality of the RepeatMasking: Example from the Macaque Genome

TE are known to cover an important part of eukaryotic genomes. As we are not interested in the behavior of repetitive elements here, it is crucial for our analysis that the genome we study have their repetitive elements carefully masked. Unfortunately, repeats differ from one species to another, and might not have been

well identified in recently sequenced genomes.

TE are also known to duplicate by bursts [40]. This happens when one repetitive element, the so-called master sequence, suddenly gets the ability to duplicate itself massively in its host genome. For this reason, all duplicates stemming from one such event have roughly the same age.

When we analyzed the MLD of the self-alignment of the Macaque genome, we obtained an $\alpha = -3$ power-law distribution (see Fig. 3.16). However, we noticed a deviation from the power-law behavior for eight lengths: the distribution exhibits several peaks in the number of matches of these lengths. Namely, the number of matches for these eight lengths was from 2 to 3 times higher than the number of matches with one more or one less nucleotide (for instance there are 1342 matches of length $r = 293$, but only 508 and 516 matches of length $r = 292$ and $r = 294$). We then retrieved the sequences associated to these matches. For each of these lengths, we found that one sequence (or in some cases two highly similar sequences, that differs on one or two bps only) was responsible of roughly two third of the matches of that length. Moreover, these eight sequences were also highly similar to each others (the smaller sequences being in almost all the cases a subsequence of the longest one) and we found that they all mapped to a single repetitive element (namely MacERV2_LTR1) identified by Han *et al.* [114] in 2007. For an unknown reason, this element has not been carefully masked in the RepeatMasked version of the Rhesus Macaque genome that can be found in the `Ensembl` database [83]. We then computed a MLD from the self-alignment of the Rhesus Macaque genome where we manually masked this element and obtained almost the same distribution without the eight peaks mentioned above (see Fig. 3.16).

3.5.4 Conclusion

In this Chapter, we have shown that different duplication processes that occur in eukaryotic genomes could shape the power-law distributions observed in the MLD

computed from the self-alignment of these genomes. We detailed how different biological mechanisms resulted in power-laws with different exponents.

MLDs can also be computed from the comparison of genomes of different species, and, fascinatingly, also exhibit power-law distributions. While it may be tempting to link these power-laws to the one observed in the self-alignment of genomes, we will show in the next chapter that they stem from a different evolutionary process, independent of any duplication mechanism.

Chapter 4

Comparative Alignment

In this Chapter, we study the properties of Match Length Distributions computed from the comparative alignment of distinct organisms.

We first show that these MLDs also exhibit a power-law distribution for a wide range of comparisons. We show that these power-laws are not linked to any duplication process, and show that the variation of substitution rate along genomes is a necessary condition of the appearance of such a power-law.

4.1 MLDs of Comparative Alignments

To compute MLDs from the comparison of genomes of two distinct organisms, as for self-alignment, we first retrieve all exact matches between the genomes, using the `mummer` software [115]. Interestingly, the MLD computed from the alignment of different eukaryotic species also results in power-law distributions, as first reported by Salerno *et al.* [97] (see also Gao and Miller [116]). We reproduce such inter-species comparisons on Figs. 4.1 and 4.2. The power-law behavior holds for the comparison of a wide range of species, although not for all comparisons. From the empirical data, we note that if species are very closely related (as exemplified on Fig. 4.1(A) by the comparison of the Human genome to the Chimpanzee genome), the MLD exhibits an exponential distribution. For the comparison of

more distantly related species (Human and Mouse, Human and Coelacanth as well as many other pairs, see Fig. 4.2), the MLD exhibits a clear power-law with exponent $\alpha = -4$. Finally, when the distance between the two species gets really large (as exemplified by the comparison of Human and Fly genomes, see Fig. 4.1 (D)), we observe few matches whose length distribution is not a power-law anymore. The goal of this Chapter is to discuss models of evolution that could explain the behavior observed for all these comparative alignments.

In this Chapter and as in previous Chapters, we refer to all changes that affect the DNA sequence of a genome, i.e point mutation, short insertion or deletion etc., as "mutations". As from now on we focus on the comparison of genomes of different species, we only consider changes that are fixed in the population of a species. In this sense, the mutation rates we discuss in the following are effective mutation rates, taking into account events that are fixed in the population only.

4.2 Pseudogene Hypothesis

We have already described in Section 3.4 a duplication mechanism that results in an $\alpha = -4$ power-law distribution. Our first working hypothesis was that the comparative $\alpha = -4$ power-laws were related to the one observed in retropseudogenes. Especially, if some genes are highly similar in both species and retroduplicating in at least one of the two species, the comparison of the two sets of retroduplicated genes is expected to produce an $\alpha = -4$ power-law.

However, when we compared the two "processed pseudogenomes" (that are both constructed from the concatenation of all reported human and mouse pseudogenes, see Section 3.4.1 for more details) of Human and Mouse, we found only a few exactly conserved sequences, and no match longer than 100 bps, as shown on Fig. 4.3. Moreover, the MLD computed from this comparison is not shaped as an $\alpha = -4$ power-law. Additionally, it has been shown that some genes were more prone to be reverse-transcribed than others [117]. Comparing the sequences of these genes in Human and Mouse, as for instance the RPL21 gene, we found

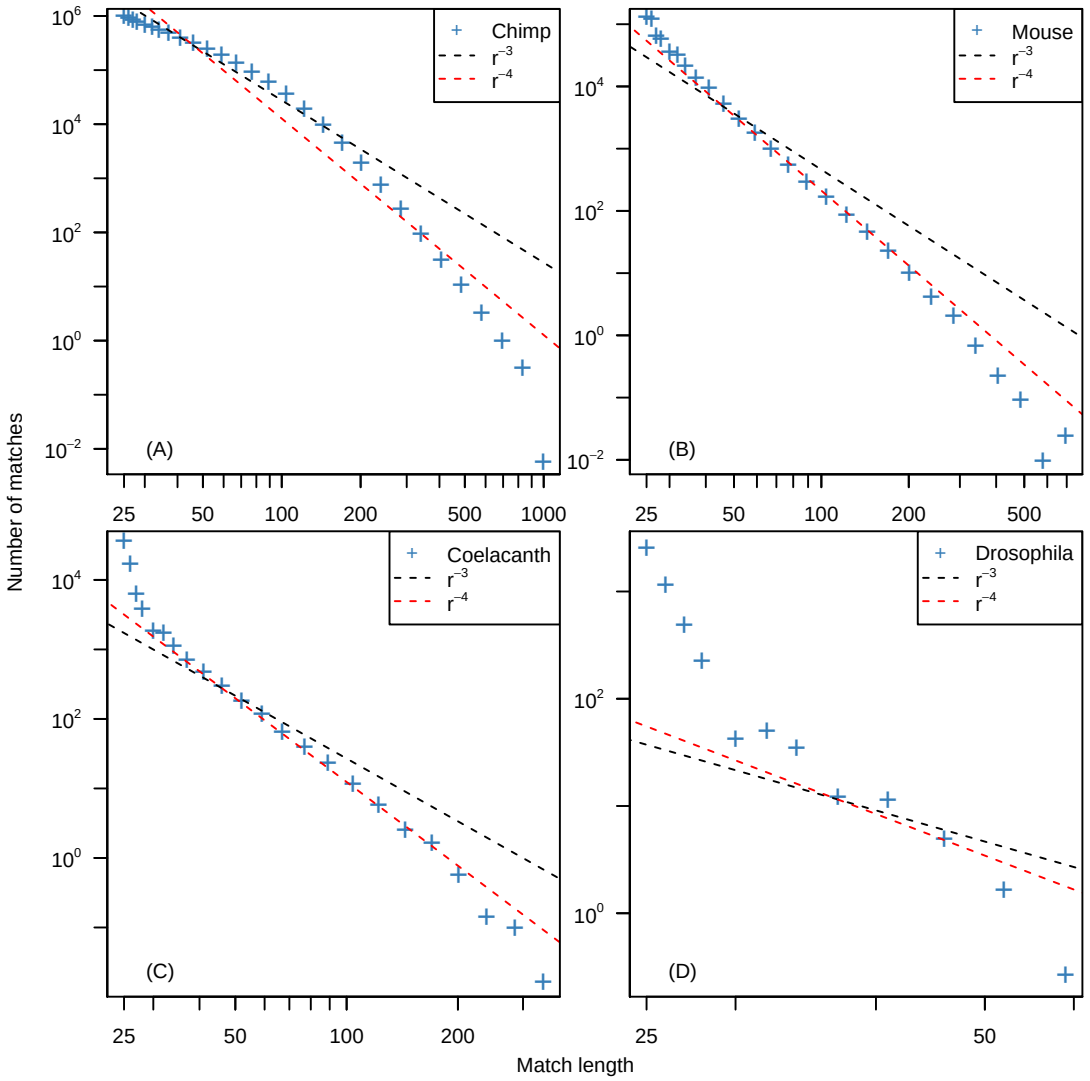

FIGURE 4.1: The MLD computed for the comparison of the human Repeat-Masked genome to different genomes. The dashed lines represent power-laws with exponent $\alpha = -3$ and $\alpha = -4$. All the empirical data are represented using logarithmic binning (see Section 2.2 for details). (A) The comparative alignment of human and chimp genomes. (B) The comparative alignment of human and mouse genomes. (C) The comparative alignment of human and coelacanth genomes. (D) The comparative alignment of human and fly genomes.

that these genes have already accumulated several independent mutations in both genomes. For instance, the longest match between the human and mouse RPL21 genes is of length 35 bps (see Fig. 4.4). For this reason, this process cannot explain the appearance of long matches in the Human/Mouse comparison and the $\alpha = -4$ power-law observed in the Human/Mouse comparative MLD. More generally, a deep analysis of processed ribosomal proteins has shown that there is

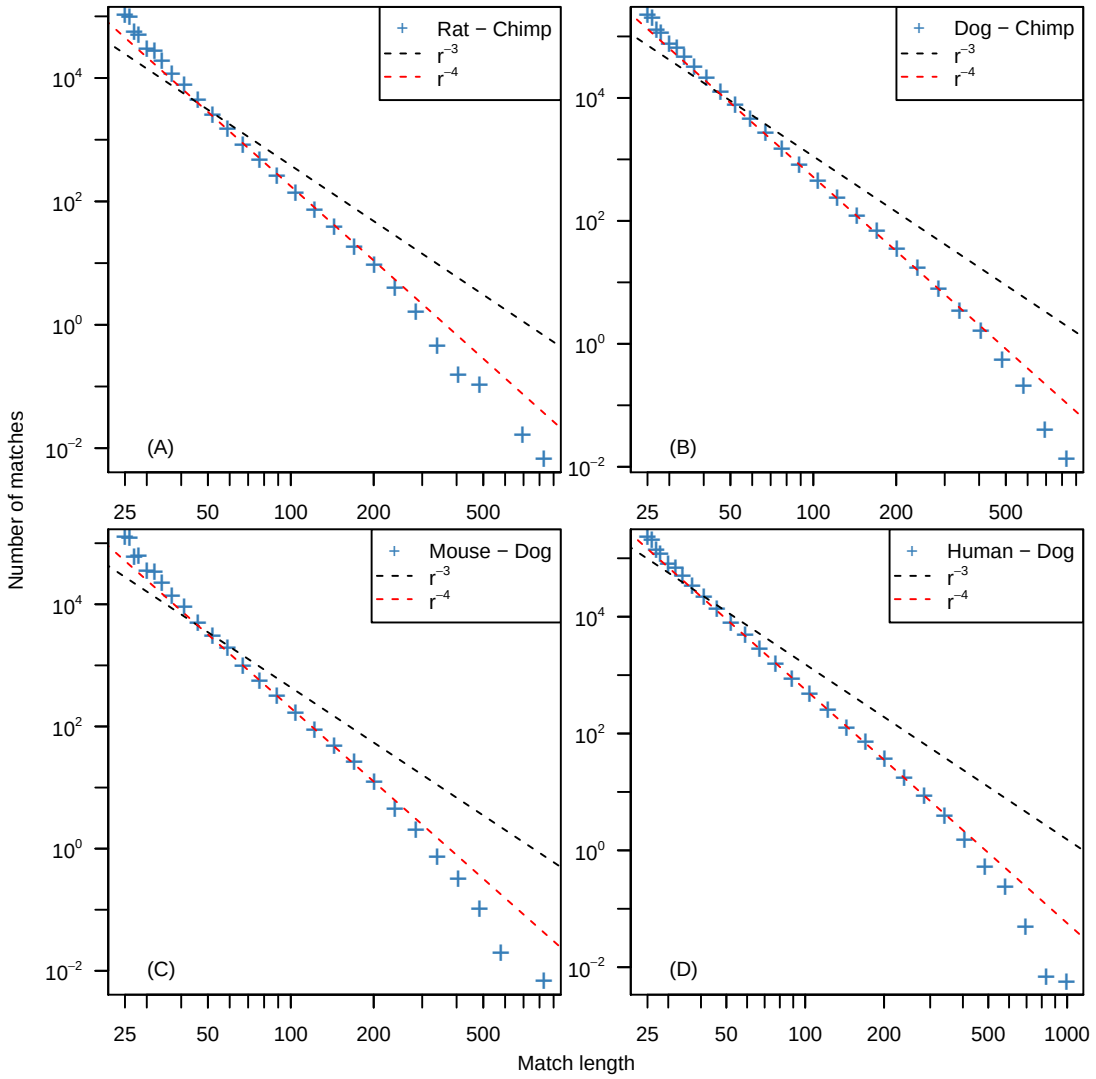

FIGURE 4.2: The MLD computed for the comparison of several pairs of RepeatMasked genomes. The dashed lines represent power-laws with exponent $\alpha = -3$ and $\alpha = -4$. All the empirical data are represented using logarithmic binning (see Section 2.2 for details). (A) The comparative alignment of rat and chimp genomes. (B) The comparative alignment of dog and chimp genomes. (C) The comparative alignment of mouse and dog genomes. (D) The comparative alignment of mouse and rat genomes

almost no preservation of these pseudogenes between rodent and the human lineage [118]. The results we obtain suggests that this statement could be extended to all pseudogenes, regardless of their parent gene.

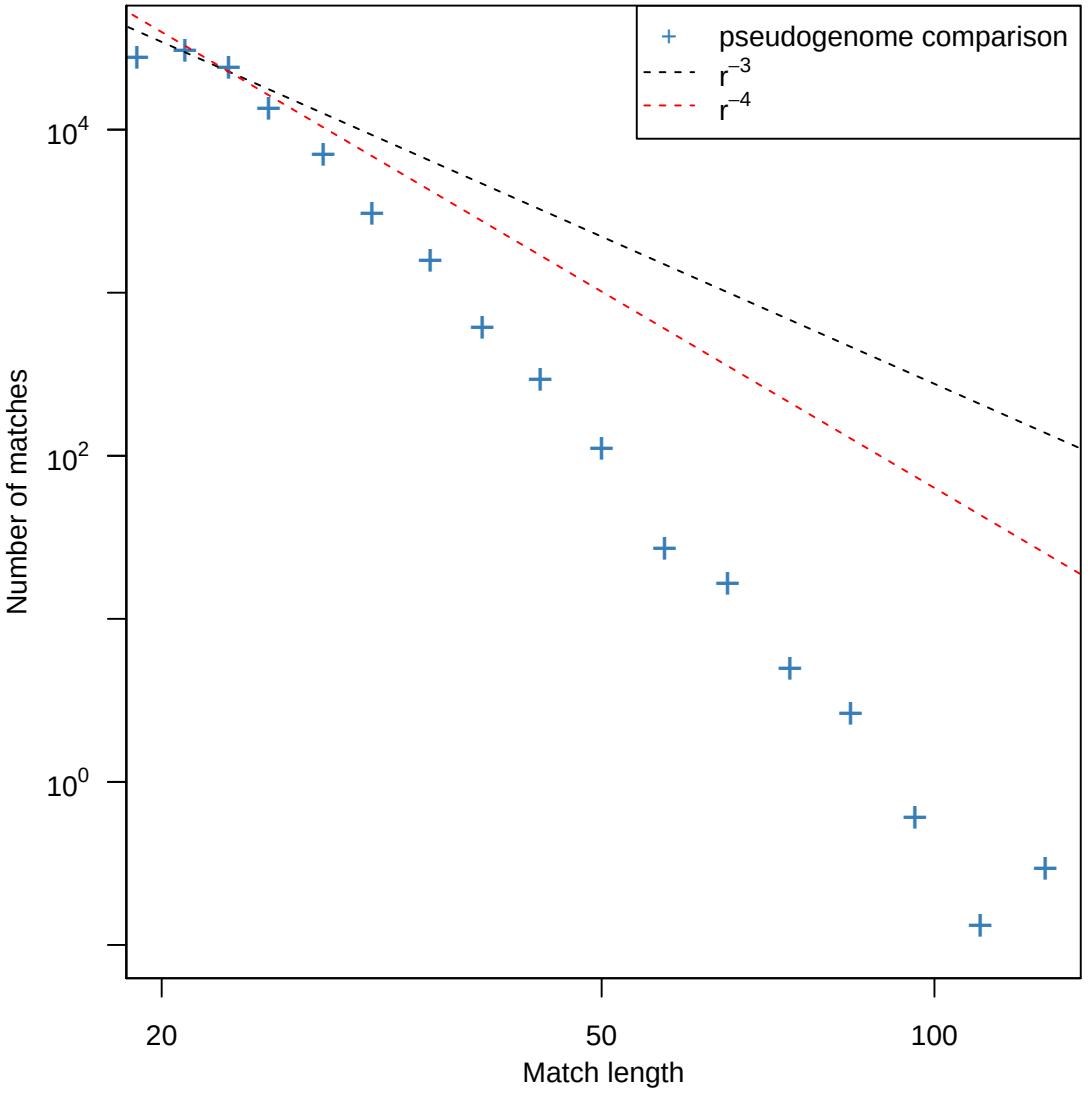

FIGURE 4.3: The MLD computed for the comparative alignment of the Human and Mouse processed pseudogenomes, constructed using the procedure of Section 3.4.1. Dashed lines represent power-law with exponent $\alpha = -3$ and $\alpha = -4$. The MLD is represented using logarithmic binning (see Section 2.2 for details).

4.3 Ladder of Trees

For evolutionary distant organisms, the existence of any long match in the comparative alignment is due to long conserved elements. Such long conserved elements have been found in vertebrates and are referred to as ultraconserved element (UCE) [119]; for a review on this topic, see Dermitzakis, Reymond, and

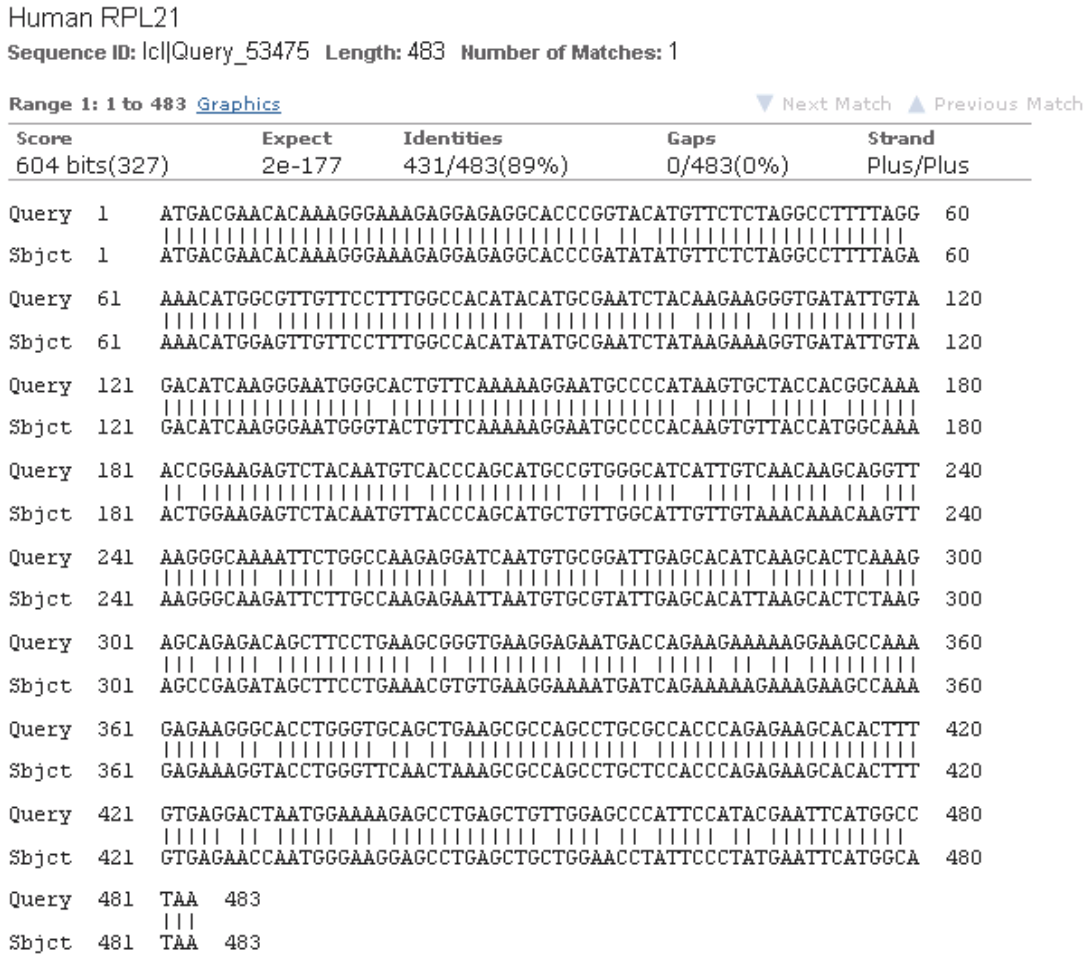

FIGURE 4.4: The screenshot of a `blast` alignment of the two transcripts of the RPL21 genes of Human and Mouse. Each transcript is 483 bps long. Although these two transcripts are very closely related, the longest exact match between the two sequences is 35 bps long only.

Antonarakis [120]. In this section, we develop a model that could explain the appearance of the $\alpha = -4$ power-law distribution through the interplay of sequence conservation and segmental duplication of conserved elements.

According to this model, long exact matches observed in comparative alignments could be the result of two different phenomena. The first class of matches would result from non mutating elements (denoted NMEs in the following), i.e. segments of the genome that have not been subject to any mutation in both species since their split. The second class would result from recent segmental duplications of a non mutating elements (denoted DoNMEs in the following). Let us assume that DoNMEs mutate with a constant mutation rate μ, and that both NMEs and

DoNMEs duplicate with the same duplication rate λ. Thus, the evolution of one DoNME family is well described by a Yule tree (see Chpater 2.3). We further assume that duplications and mutations occur independently in both species. In this case, DoNMEs are clustered into families, and each family is linked to one particular NME. We show the evolution of one DoNME family and its corresponding NME in the two species as a tree on Fig. 4.5.

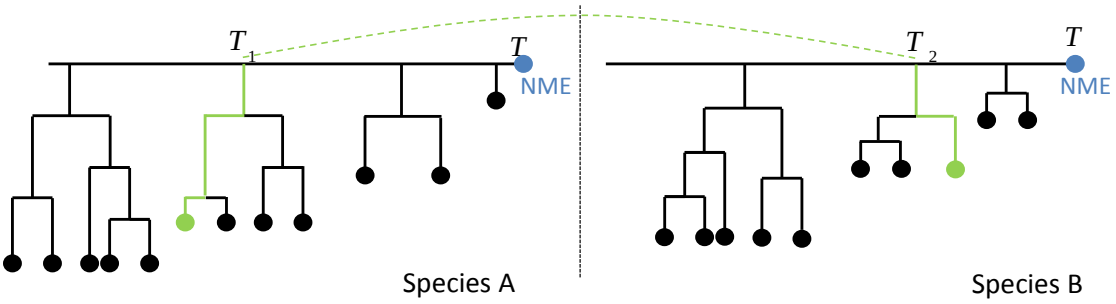

FIGURE 4.5: The evolution of a DoNME family in two distinct genomes (left and right Trees). Each black circles represent a member of the DoNME family, and blue circle represent the same NME in the two different genomes. The vertical dimension represent the evolutionary time, and the horizontal dimension is meaningless. As NMEs do not mutate, they do not move in the vertical dimension. The green path highlight the evolutionary distance between two DoNMEs that stem from the duplication of the same NME, but in two different genomes.

As each DoNME family can be described by a Yule tree, from Eq. (2.9) we know that there are $e^{\lambda(T-T_1)}$ leaves in each DoNME family, with T_1 being the time of the pioneering duplication event. Then, if the total evolution time since the split is equal to T, one gets that the average density of pairs separated by an evolutionary distance τ is given by

$$N(\tau|T) = \int_0^T dT_1 dT_2 e^{\lambda(T-T_1)} e^{\lambda(T-T_2)} \delta(\tau - \mu[(T-T_1)+(T-T_2)]), \tag{4.1}$$

where T_1 and T_2 represent the time of the duplication event in each genome. After integration over T_1 and T_2 one gets

$$N(\tau|T) = \begin{cases} \dfrac{\lambda^2 e^{\frac{\lambda\tau}{\mu}}}{\mu^2}\tau & \text{for } 0 \leq \tau \leq T\mu \\ \dfrac{\lambda^2 e^{\frac{\lambda\tau}{\mu}}}{\mu^2}(2T\mu - \tau) & \text{for } T\mu \leq \tau \leq 2T\mu. \end{cases} \tag{4.2}$$

Combining this equation with Eq. (3.6) for T sufficiently large, we find :

$$M(r) = \frac{6\lambda^2}{\mu^2} \frac{K - \frac{\lambda}{\mu}}{\left(r - \frac{\lambda}{\mu}\right)^4}. \tag{4.3}$$

Therefore, assuming that the MLD for the comparative alignment of two distant species mainly results from DoNMEs and NMEs, it is expected to follow Eq. (4.3), that is an $\alpha = -4$ power-law.

To verify our hypothesis, we reconstructed the different DoNME families. To do so, we first created a library containing the sequences of all exact matches between the two species. Then, we compared each of these sequences to the other sequences of the library. Whenever two sequences shared one or more exact matches longer than 20 bps, we grouped the two sequences in the same family. If one or both sequences already belonged to a family, we merged the two families. Note that this way, we may miss old duplication events (if the two duplicated sequences have already highly diverged). As we want to explain the appearance of a power-law for matches longer than 20 bps, these old events cannot be responsible for the statistical property we want to explain.

We then computed the MLD of each set to obtain one MLD for unique matches (we filtered out all matches that belong to a family of size larger than one) and one MLD for non-unique matches (we filtered out all unique matches). The MLD containing duplicated matches exhibits the expected behavior, but surprisingly, the MLD constructed only with unique matches also exhibits a power-law with exponent $\alpha = -4$ (see Fig. 4.6). Moreover, more than two third of the matches are unique and are not linked to any other match. This indicates that unique matches quantitatively dominate the comparative MLD, and that the power-law observed in comparative alignment does not stem from a duplication process. This disproved our working hypothesis and leads us to the next model where all matches are unique.

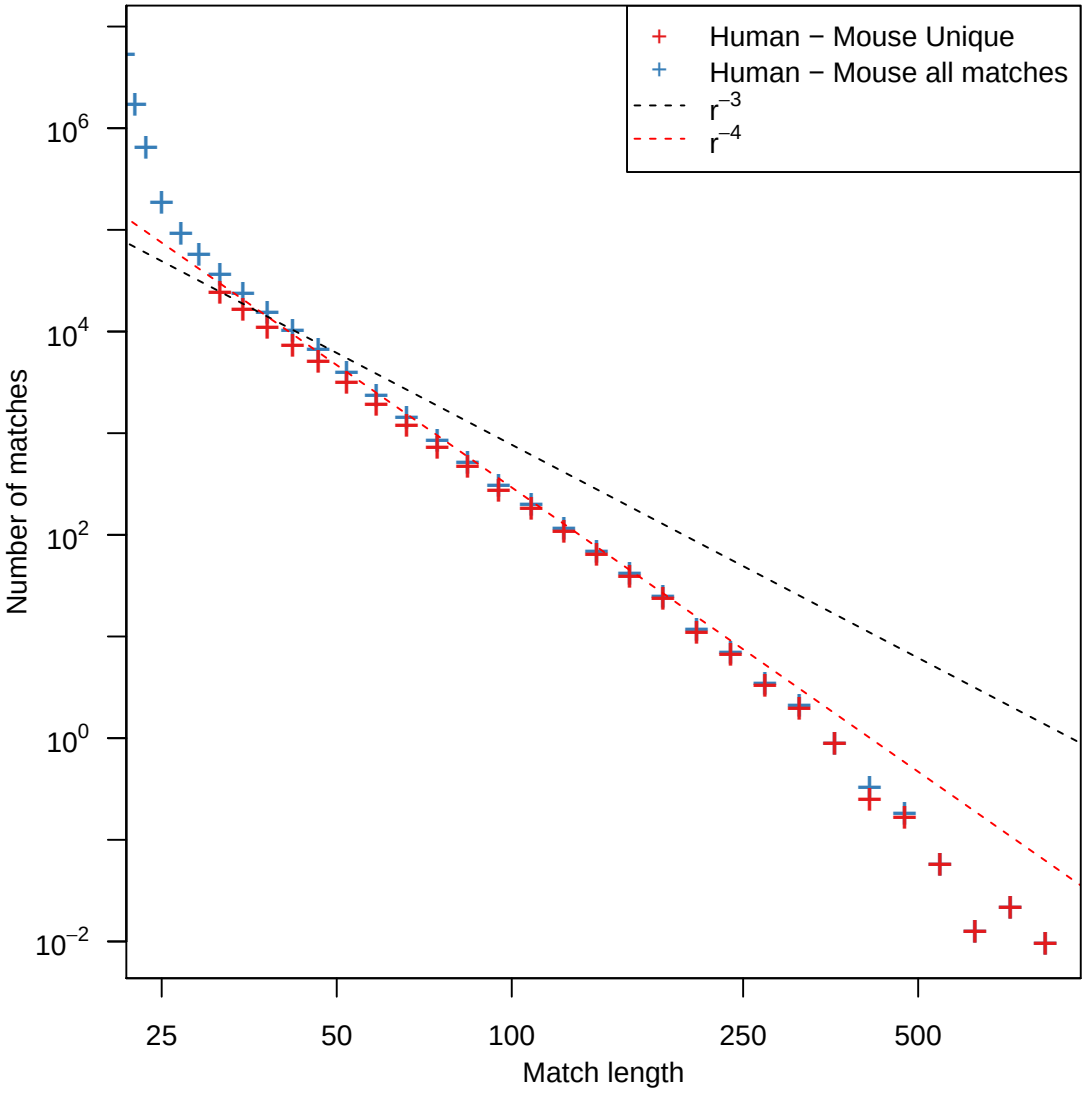

FIGURE 4.6: The MLD computed from the comparative alignment of the human and mouse genomes. Here, we filtered out all the matches that share similarity (i.e. at least one exact match longer than 20 bps) with another match of the comparative alignment. To do so, we first retrieved the sequences of all the matches obtained from the comparative alignment of the human and mouse genomes. We then compared each of these sequences to all the other sequences. Any similarity found between the two sequences indicates that a part of this sequence is present in more than one copy in at least one of this two genomes, and we filtered it out of our analysis. Doing so, we removed approximately one third of the matches obtained in the comparative alignment. The MLD after filtering still exhibits a -4 power-law distribution.

4.4 The Evolution of Conserved Regions

4.4.1 Theoretical Calculations

Comparing Species Shortly after the Split — Shortly after a speciation event, the genomes of the two resulting species, denoted by A and B, are almost identical.

As a consequence, the alignment of the two genomes exhibits many long and exact matches, which are either orthologs (along the main diagonal of the alignment grid) or paralogs (off diagonal matches on the alignment grid). The latter are the reminiscences of segmental duplication in the genome of the common ancestor of A and B, and are quantitatively less important than orthologous matches (see previous section). The MLD obtained when comparing these two genomes has always an exponential tail, which stems from orthologous matches. For short evolutionary times we can assume that mutations happen at random positions along the two genomes and, therefore, the MLD is qualitatively described by the stick breaking model where the initial stick length (now denoted by C) is the length of the alignable orthologous part of the two genomes. In that sense, what we observe is the result of one stick breaking experiment where the length of the stick is the length of an entire genome. The tail of such a MLD is therefore exponentially distributed and follows (see Section 3.1.2):

$$m(r,\tau) = \left[2\tau + \tau^2 (C - r)\right] \exp(-\tau r). \tag{4.4}$$

Indeed, an exponential distribution is observed in empirical data, for instance for a Human-Chimp comparison (see Fig. 4.1(A) and [116]).

The comparison of Distantly Related Species — Following the latter simple process, the number of long matches decreases fast with the divergence between the two species. For this reason, this process alone would not result in long matches in an alignment of genomes of highly divergent species, like, for instance, Human and Mouse. As the divergence between Human and Mouse is of the order of 25% [121], according to the stick breaking model, the length of the longest expected match between these two genomes would be $r = 72$ bps (if we assume that both RepeatMasked genomes are of length 1 Gbp), see Eq. (4.4). However, when comparing Human and Mouse, we obtained 820 exact matches of length 72 bps, and the length of the longest exact match is $r = 781$ bps. Moreover, the MLD observed for Human-Mouse alignment exhibits a heavy tail, shaped as a power-law with an exponent $\alpha = -4$ (see Fig. 4.6). This distribution stems from the many

well conserved regions that are shared by the Human and the Mouse. In total, we obtained more than $6 \cdot 10^5$ exact matches longer than 25 bps, and all together, they span more than 22 Mbps.

If we assume that such a high degree of conservation is the consequence of some biological functionality, it follows that there are regions that evolve at their own (slow) speed, i.e. with a lower mutation rate (see Fig. 4.7). As each such region can play a different role in the two considered genomes, the mutation rate may be different for the same region in the two different genomes. This leads us to assume that the evolutionary distances between orthologous regions are not constant, but are drawn from some distribution. In the following, we assume that each genome is organized in regions of mean size C, that the mutation rate in each region is constant, and that the different values of the mutation rate along the different regions are drawn from some continuous distribution. We demonstrate that this assumption leads to a qualitative change in the shape of the MLD.

Calculating $N(\tau)$, the Number of Regions at a Distance τ from each others — The evolutionary distance between a pair of orthologous sequences is given by

$$\tau = \tau_A + \tau_B, \tag{4.5}$$

where τ_A is the evolutionary distance separating a region in A to its orthologous region in the last common ancestor of A and B, likewise for τ_B (see Fig. 4.7 for an illustration). If we assume that the mutation rates are independent in each genome, for a given evolutionary distance τ, the two distances τ_A and τ_B can take different values, still satisfying Eq. (4.5). Namely, we have:

$$P(\tau_A + \tau_B = \tau) = \int_0^\tau P(\tau_A = x) P(\tau_B = \tau - x)\, dx, \tag{4.6}$$

and the number of sequence regions separated by the evolutionary distance τ is therefore given by:

$$N(\tau) = \int_0^\tau N_A(\tau - \tau_B) N_B(\tau_B)\, d\tau_B, \tag{4.7}$$

where $N_A(\tau)$ is the number of sequences in species A separated by the evolutionary distance τ from its orthologous sequence in the last common ancestor of A and B, likewise for $N_B(\tau)$ (see Fig. 4.7). Since the divergence τ_{A_i} between each sequence i of species A and its homologous sequence in the common ancestor of A and B is simply given by:

$$\tau_{A_i} = \mu_{A_i} T. \tag{4.8}$$

thus, $N_A(\tau)$ is directly proportional to the number of sequences whose mutation rate is $\mu_{A_i} = \tau/T$, that is, to the distribution of mutation rate in genome A. For simplicity in the following, we refer equivalently to N_A (resp. N_B) or to the distribution of mutation rates in genome A (resp. B).

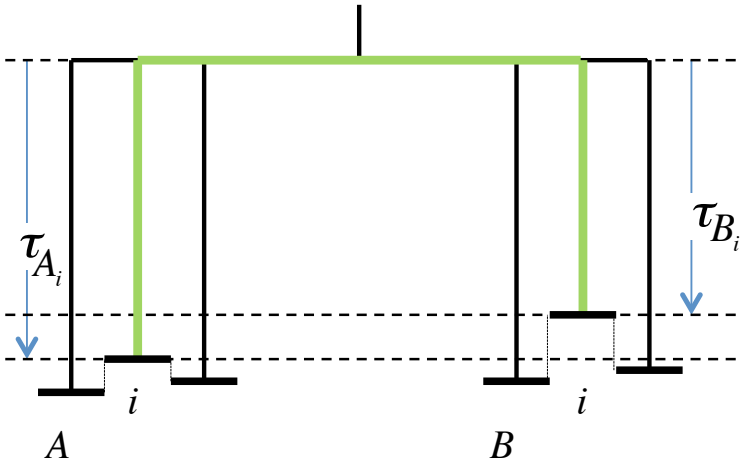

FIGURE 4.7: An example of the evolution of two divergent genomes. Different regions of the two species A and B evolve with different rates. The evolutionary distance separating two orthologous regions i (green path) is the sum of the evolutionary distance covered by this genomic region in both species since their split.

In general, following Eqs. (3.2) and (3.6), and replacing the mean length of a duplication K by the mean length of a conserved region C, the MLD is given by

$$M(r) = \int_0^\infty \left[2\tau + \tau^2 (C - r)\right] \exp(-\tau r) N(\tau) \, d\tau \tag{4.9}$$

for $r < C$. Long matches correspond to sequences at small evolutionary distances τ. Thus, the distribution $M(.)$ for long length (i.e. $r \gg 1$) is controlled by the

integration over small values of τ in Eq. (4.9). For such small values of τ the function $N(\tau)$ can be expanded in a Taylor series next to $\tau = 0$:

$$N(\tau) = N(0) + \left.\frac{dN(\tau)}{d\tau}\right|_{\tau=0} \tau + \mathcal{O}\left(\tau^2\right). \tag{4.10}$$

Eq. (4.7) implies that $N(0)$ always vanishes, such that the next term, $N'(0)\tau$ linear in τ, becomes dominant. It follows that:

$$\begin{aligned} M\left(r\right) &\simeq \int_0^\infty \left[2\tau + \tau^2\left(C - r\right)\right] \exp\left(-\tau r\right)\, \tau\, N'(0)\, d\tau \\ &= \int_0^\infty 2\tau^2 \exp\left(-\tau r\right) N'(0) d\tau + \int_0^\infty \tau^3 \left(C - r\right) \exp\left(-\tau r\right) N'(0)\, d\tau \end{aligned} \tag{4.11}$$

Applying the integration by parts technique to the second term gives:

$$\begin{aligned} M(r) &= N'(0) \int_0^\infty 2\tau^2 \exp(-\tau r)\, d\tau + N'(0) \left[0 + \frac{3(C-r)}{r} \int_0^\infty \tau^2 \exp(-\tau r)\, d\tau\right] \\ &= N'(0)\, \frac{3C - r}{r} \int_0^\infty \tau^2 \exp(-\tau r)\, d\tau. \end{aligned} \tag{4.12}$$

Integrating by parts again finally leads to:

$$M(r) = N'(0) \frac{6C - 2r}{r^4}, \tag{4.13}$$

in the regime $1 \ll r < C$, in agreement with the observed MLD between distantly related genomes. It follows that the match length distribution exhibits an $\alpha = -4$ power-law unless the first derivative $dN(\tau)/d\tau|_{\tau=0}$ also vanishes. We can apply Leibniz rule of derivation (see Flanders [122]) to Eq. (4.7) to calculate the value of the first derivative, which results in:

$$\frac{dN(\tau)}{d\tau} = N_A(0) N_B(\tau) + \int_0^\tau N'_A(\tau - \tau_B) N_B(\tau_B) d\tau_B, \tag{4.14}$$

and thus

$$\left.\frac{dN(\tau)}{d\tau}\right|_{\tau=0} = N_A(0) N_B(0). \tag{4.15}$$

The pre-factor of the MLD in this case depends both on C and on the value of $N_A(0)$ and $N_B(0)$, which are the number of regions which have not mutated since the split of species A and B.

4.4.2 Simulations

In order to illustrate our results, we simulated a model that belongs to the class where $dN(\tau)/d\tau|_{\tau=0} \neq 0$. For these simulations, we let a synthetic genome evolve according to two simple processes, point mutations and segmental duplications. The duplication process is similar to the one described in the simple model of Section 3.2.2. As previously, the duplication rate is denoted by λ and the duplication length is constant equal to K.

For the mutation process, the genomes are first divided into small regions of constant length C. For each region, we independently draw a mutation rate from an exponential distribution with mean 1. We chose the exponential distribution because it is the distribution that minimizes a priori information when only its average is known, but we obtained equivalent results using a uniform distribution, in agreement with the results of our theoretical results showing that only the value of the mutation rate distribution in 0 affects the MLD. For each region, the mutation rate is constant in time all over the simulation.

As in the previous cases, we start with a random iid sequence S_0. We first let the sequence evolve for an initiation time t_0 to obtain the ancestor sequence S_I. We then model a divergence event. We copy S_I to obtain two sequences, S_A and S_B, and redraw independently from the same distribution the mutation rates for each region of each species. We then let both sequences evolve independently according to our model for the same time t_1. For more details on the simulation procedure, see Section 2.4.

In Fig. 4.8, we present the MLD computed from simulated sequences for a self-alignment (equivalent to divergence time $t_1 = 0$), and for different divergence times $t_1 = 0.01$, $t_1 = 0.2$ and $t_1 = 5$. Qualitatively, these simulations exhibit

the same behavior as the self-alignment of the human genome, the comparison between human and chimpanzee genomes (for $t_1 = 0.01$), the comparison between human and mouse genomes (for $t_1 = 0.2$) and the comparison between the human and fruit-fly genomes (for $t_1 = 5$), respectively, see Fig 4.1 panel (B), (C) and (D).

Note that the mutation rate in this model is constant over small regions of length of the order of the longest expected match (in the simulations we presented, $C = 1000$). In the extreme case where the mutation rate is independently chosen for each base pair, i.e. with regions of length $C = 1$, the power-law behavior is never observed, and the MLD shift directly from an exponential distribution for closely related species to no match for distantly related genomes (data not shown).

4.4.3 Discussion

In this Chapter we have shown that only certain evolutionary scenarios are able to account for empirical power-law behaviors in the MLDs of the comparative alignment of two distantly related genomes. The only evolutionary process involved in these scenarios is point mutations, and the power-law distribution results from the fact that the mutation rate is not constant all over the genome. This reflects the existence of neutrally evolving regions (fast evolving regions) and well conserved regions of the genomes, as for instance ultra-conserved elements (slow evolving regions).

The Distribution of Mutation Rates —

To obtain a power-law in the comparative MLD, it is essential that the mutation rate is continuously distributed. For instance, if all regions of the genome have the same mutation rate, then $N_A(\tau)$ (resp. $N_B(\tau)$) is zero for all value of $\tau \neq \mu t$ and thus $N(\tau) = I_{\tau=2\mu t}\Lambda$, where Λ is a constant. In this case we cannot write the Taylor expansion of Eq. (4.10), and it follows from Eq. (4.9) that the MLD is a simple exponential distribution. Similarly, a model that would only consider two classes of regions – well conserved and fastly evolving – would not result in this

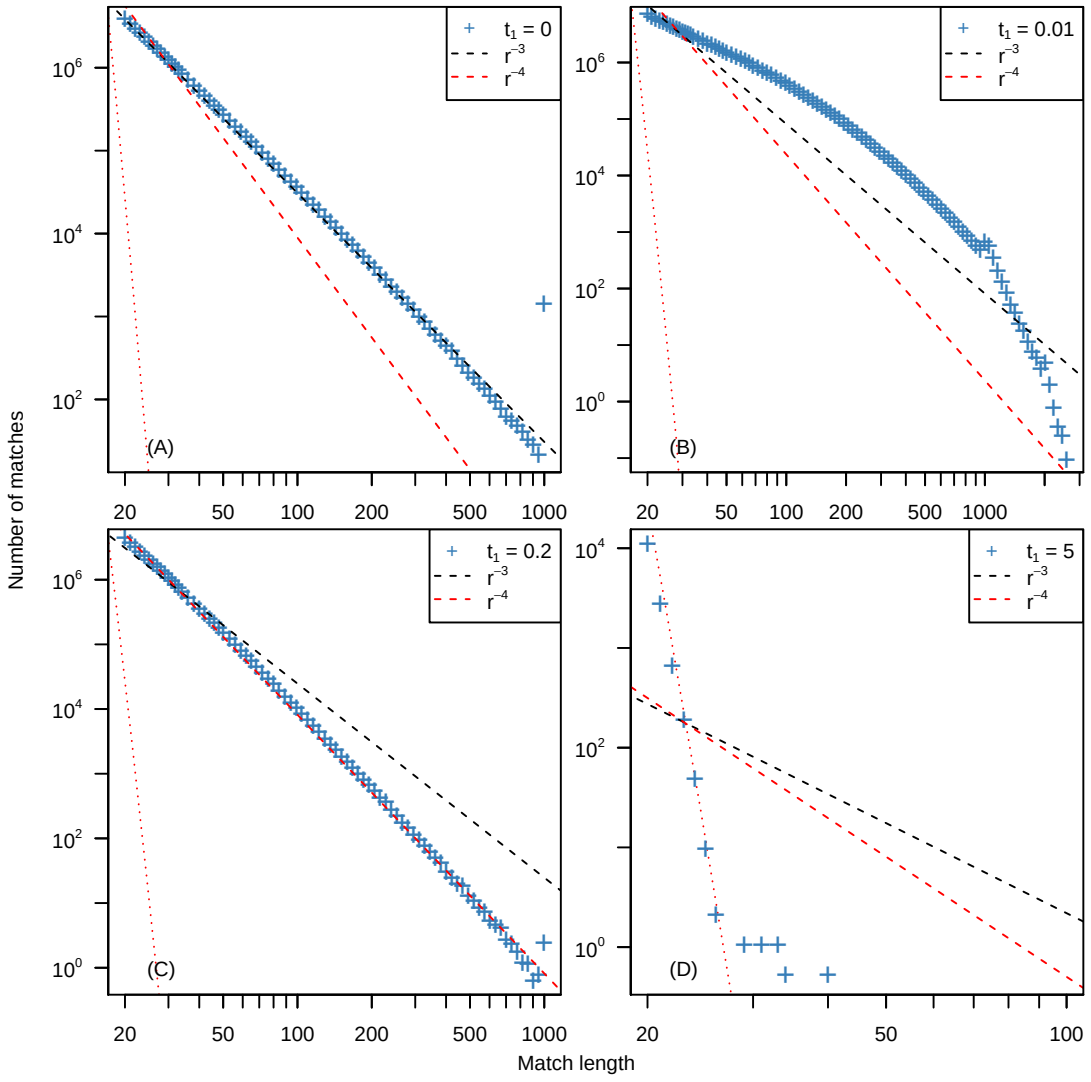

FIGURE 4.8: The match length distribution computed for simulated sequences with various divergence times. In all panels, the red dotted line represents the theoretical distribution obtained when computing the same experiment on random iid sequences with the same length and the same nucleotide frequencies than the simulated sequences. The dashed lines represent power-law functions proportional to $1/r^3$ (black) and $1/r^4$ (red), where r is the match length. All empirical data are represented using logarithmic binning. Each plot shows the histogramm obtained for 10^4 sequences of length 10^6 bps. For all simulations, the duplication rate per bp $\lambda = 10^{-3}$, the length of a duplication $K = 1000$ bps and the length of a mutating region $C = 1000$ bps. (A) The self-alignment of the common ancestor after $t_1 = 0$. (B,C,D) The comparative alignment of two sequences with divergence time $t_1 = 0.01$ (B), $t_1 = 0.2$ (C) and $t_1 = 5$ (D).

$\alpha = -4$ power-law. To sum up, to obtain a power-law tail with $\alpha = -4$ up to a certain length $r_{\max}$ in the MLD computed from the comparative alignment of two genomes requires three conditions on the mutation rates in both species:

– The mutation rate is constant inside well conserved DNA regions of mean length C, with $C > r_{\max}$.

– The distribution of mutation rate across well conserved regions is smooth (at least differentiable) in both species, such that we can write the Taylor expansion of $N(\tau)$.

– The distributions of mutation rate in both species, N_A and N_B, do not vanish at zero, meaning that it exists, in each genome, non-mutating regions. Importantly, these non-mutating regions do not need to be orthologous.

These conditions are quite general and can be fulfilled by a wide range of models [123]. Therefore, the observation of a MLD exhibiting an $\alpha = -4$ power-law alone does not allow to decide which of these models describes best the actual biological mechanisms responsible for the mutation rate variation.

The role played by C, the mean length of well conserved regions, in comparative alignment, is equivalent to the role played by K, the mean length of segmental duplications, in the case of self-alignment such that the discussion on the distribution of K in Section 3.2.1 on page 62, also apply to C. Thus, if C is not constant, it does not affect the power-law shape of the MLD, as long as $C \gg r_{\max}$. Note that the prefactor of the MLD here depends on C and on the values of $N_B(0)$ and $N_B(0)$, namely the number of regions which have not mutated yet in the species A and species B genomes. As all these 3 parameters are unknown, we cannot use our analysis to estimate any of them. However, the length up to which the power-law behavior holds gives a lower bound for the value of C. For example, the MLD computed from the Human-Mouse comparison exhibits a power-law behavior up to a length of $r = 300$ bps (see Fig. 4.6), indicating that $C > 300$ bps in both genomes.

Note that there is an apparent paradox in the fact that the prefactor of the MLD is higher for the comparison of the Human and Dog genome (see Figs. 4.2 (D) and 4.1 (B)) show) than for the comparison between Human and Mouse, while Human and Mouse are more closely related than are Human and Dog (the split

between Human and Dog occurred after the split between Human and rodents). Still, the mutation rate in rodent genomes is known to be roughly 3 times higher than in both Human and Dog genomes ([124]). Thus, the absolute evolutionary distance between Human and Dog is smaller than between Human and Mouse, which explains why we find more matches when comparing Human and Dog than when we compare Human and Mouse. A similar behavior can be observed for the triplet of species Rat-Dog-Chimpanzee (see Fig. 4.2, panel (A) and (B)).

The case of Paralogs — As mentioned in section 4.3, in the comparative alignment of any two species, the majority of matches are unique. In that sense, these matches "dominate" the distribution. One can artificially remove these unique matches from the alignment. The remaining paralogous (off-diagonal) DNA segments are expected to exhibit an $\alpha = -3$ power-law for closely related species, because in this case the comparative alignment is similar to the self-alignment of one of the species. However, as the divergence between the two species increases, the value of $N(\tau)$ close to zero decreases and vanishes. Thus, the $\alpha = -3$ power-law is expected to slowly switch to an $\alpha = -4$ power-law, similarly to the MLD of unique sequences. Such a trend was observed recently [116].

Power-laws in MLDs of other Comparisons — In Section 3.4, we described an other $\alpha = -4$ power-law, that stemmed from the self-comparison of segments duplicated via reverse-transcription of mRNA molecules. In that case, we were able to calculate the exact functional form of $N(\tau)$, namely:

$$N(\tau) = \begin{cases} \dfrac{\lambda^2 K^2}{2\mu^2} \dfrac{1}{1+a}\, \tau & \text{for } 0 \leq \tau \leq (1+a)\mu T \\ \dfrac{\lambda^2 K^2}{2\mu^2} \dfrac{-1}{1-a} (\tau - 2\mu T) & \text{for } (1+a)\mu T \leq \tau \leq 2\mu T. \end{cases} \tag{4.16}$$

Interestingly, in this case as well, $N(0) = 0$ and $dN(\tau)/d\tau|_{\tau=0} > 0$. More generally, if $N(\tau)$ scales as τ^β (with $\beta \in \mathbb{N}$) for small values of τ, the exponent of the expected power-law is $\alpha = -(3+\beta)$. Therefore, different integer power-laws could be observed if different derivatives in the Taylor expansion of $N(\tau)$ vanish. For

example, we compared the human and mouse exomes. The resulting MLD exhibits a power-law tail with an exponent $\alpha = -5$ suggesting that in this case, the first derivative $N'(0) = N_{\mathrm{H}}(0)N_{\mathrm{M}}(0)$ vanishes. The analysis of this distribution is the subject of Chapter 6. This result also indicates that the comparative $\alpha = -4$ power-law is not a feature of coding sequences. Indeed, the MLD computed from the comparison of the non coding part of Human and Mouse also exhibits an $\alpha = -4$ power-law.

In conclusion, we have shown that the distribution of match lengths in a genomic alignment of two species goes through qualitatively different regimes as the genomes diverge (Fig. 4.1). Notably, the distribution of the mutation rate along the genomes of the two species generates a power-law distribution with an exponent $\alpha = -4$ in the distribution of exact matches. Such a power-law therefore occurs naturally in the MLD of two diverging genomes and is a signature of differences in functional constraints and it is therefore not occurring neutrally.

Chapter 5

At the Crossing Between Self and Comparative Alignments: The Case of Whole Genome Duplication

In Chapter 3, we have studied duplication events that occur frequently and at small scale compared to the genome size. However, genomes are also subject to large scale duplications, that range from single chromosome duplication to whole genome duplications. In this Chapter, we study the MLD computed from the self-alignment of a genome in the case of WGD, where the entire genetic information of an organism is duplicated. We show that depending on the time elapsed since the WGD event, this MLD can be assimilated either to a comparative MLD or to a self-alignment MLD.

5.1 The Fate of a Genome after a Whole Genome Duplication

In the following, let us denote by genome I the genome just before the WGD. Immediately after the WGD event, the genome consists of two identical sub-genomes, sub-genome A and sub-genome B, both identical to genome I. We assume that both sub-genomes evolve independently after the WGD event, as would two distinct species.

We present four schematic dot plots describing the fate of a Genome after a WGD event at four different time point on figure 5.1. Just after the WGD (shown on panel (A) of Fig. 5.1), sub-genome A and B are exactly identical. It follows that qualitatively, the MLD obtained from the self-alignment of this genome is similar to the one obtained from the self-alignment of the sequence before its duplication, the only difference being that the prefactor is 4 times higher after the WGD. Notably, we obtain two very long matches corresponding to the match of the entire sub-genome A against the entire sub-genome B. We will refer to these matches as matches of the "second diagonal". Then, each of the two sub-genomes will start to accumulate mutations independently, breaking exact similarities between the two sub-genomes. As a consequence, the length of the matches stemming from ancient segmental duplications will decrease. At the same time, the breaking down of the second diagonal matches will result in several very long matches (Fig. 5.1, panel (B)). As mutations accumulates, the second diagonal will be broken in smaller pieces resulting in more and more matches, while the similarities resulting from ancient segmental duplications will dramatically decrease (Fig. 5.1, panel (C)). Finally, the two sub-genomes will reach a point where similarities due to the WGD have completely disappeared, and where the MLD results only of segmental duplications that occurred after the WGD (Fig. 5.1, panel (D)).

The MLD computed from the self-alignment of the entire genome after the WGD can be separated in 2 parts: the first part results from the self-comparison of sub-genome A and the self-comparison of sub-genome B (squares (2) and (3) on the

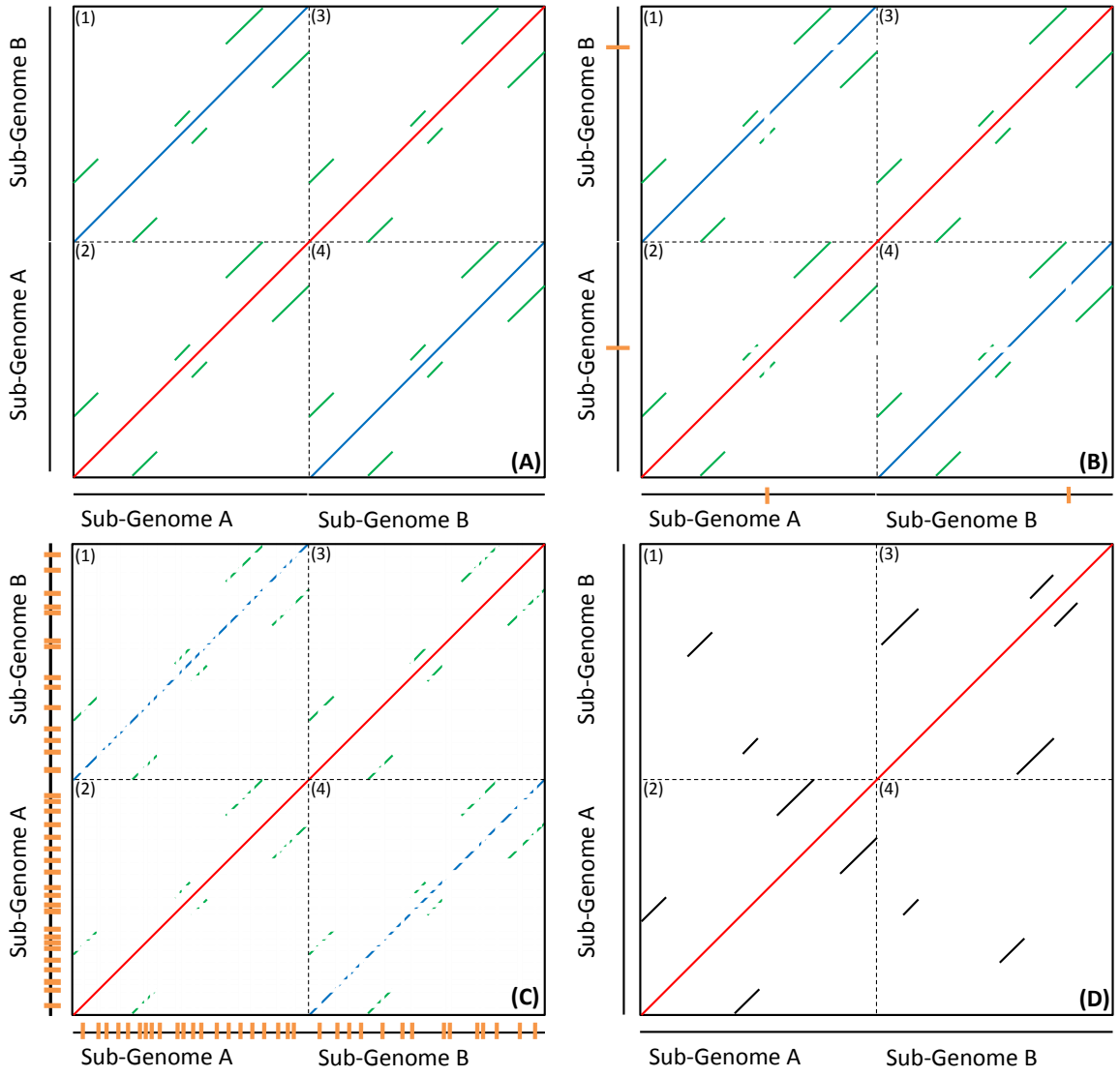

FIGURE 5.1: Schematic dot plot representation of the evolution of exact similarities after the occurrence of a WGD. Sub-genome A and sub-genome B denote the two sub-sequences that result from the WGD. Any line of the dot plot grid represent an exact match. On the dot plot grid, we represent segmental duplications that occured before the WGD in green, the trivial match of the whole genome against itself in red (first diagonal), the matches stemming from the second diagonal in blue and post WGD segmental duplications in black. Orange lines depict point mutations. (A) Just after the WGD, the two sub-genomes A and B are identical. (B) Shortly after the WGD, each sub-genome begins to accumulate independent point mutations. (C) After a longer evolutionary time, remains from ancient segmental duplications have almost disappeared, while the second diagonal give rise to a lot of matches. (D) Long after the WGD. There remains nothing from the old WGD and similarities only stem from recent SDs.

self-alignment grids of Fig. 5.1), and the second part from twice the comparison of A and B (namely the comparison of A to B (square (1) of Fig. 5.1) plus the comparison of B to A (square (4) of Fig. 5.1)). If the two sub-genomes evolve

independently, both self-alignments of A and B are expected to result in an $\alpha = -3$ power-law, similarly to what we observe for single species self-alignment. On the other hand, the comparison of A and B behaves as a comparative alignment of two divergent species, and thus, its MLD results in an $\alpha = -4$ power-law. For this reason, the MLD obtained when self aligning a genome after a WGD event is a mix of the two previously discussed cases.

Depending on the time elapsed since the last WGD event, these two processes will have different quantitative importance. If the WGD event occurred recently, then matches obtained from the comparison of A and B will dominate, and we expect an $\alpha = -4$ power-law. However, if the WGD occurred a long time ago, the evolutionary distance separating the two sub-genomes will be high, and the comparison of A and B will result in few matches. In that case, the dominating signal stems from segmental duplications that occurred in any of the two sub-genomes after the WGD. Thus, the MLD is expected to exhibit an $\alpha = -3$ power-law, like in the case of a simple self-alignment.

For this reason, after a whole genome duplication, the MLD computed from a self-alignment will result in a power-law with an exponent α equal to either -3 or -4, depending on the time elapsed since the WGD. We define t_c as the time point where the transition between the two regimes occurs. In the following, we want to calculate t_c given the mutation and segmental duplication rates.

5.2 The Transition Between the Two Regimes

As we have seen, we expect to observe a transition between two power-law following a WGD event. In this section, we try to calculate the time t_c at which this transition occur. Note that this time is a decreasing function of the length r_c one considers.

Let us assume that, like in the model of section 4.4.1, the mutation rate is constant in regions of length C on average and that the WGD occurred at time $t = 0$. Let us

also assume that the mutation rate per bp in region i, μ_i, is distributed according to an exponential distribution of mean μ. If L is the total size of the genome after the WGD, there are

$$R = \frac{L}{2C} \tag{5.1}$$

regions in each sub-genome.

If we further assume that segmental duplications still occur in the genome after the WGD event, that K is the mean length of a segmental duplication, and that the duplication rate per bp is λ, following Eq. (4.13) and Eq. (3.12), the MLD obtained from the self-alignment of one genome after a WGD event is then:

$$M(r) = \frac{\lambda K L}{\mu}\frac{1}{r^3} + 2\left.\frac{dN(\tau)}{d\tau}\right|_{\tau=0}\frac{6C-2r}{r^4}. \tag{5.2}$$

Assuming $C \gg r$, it follows that

$$M(r) = \frac{\lambda K L}{\mu}\frac{1}{r^3} + 2\left.\frac{dN(\tau)}{d\tau}\right|_{\tau=0}\frac{6C}{r^4}. \tag{5.3}$$

From Eq. (4.15) we also know that

$$\left.\frac{dN(\tau)}{d\tau}\right|_{\tau=0} = N_A(0)N_B(0). \tag{5.4}$$

If we assume that mutations occur independently inside each region, the first mutation occurs on average in each region i after time $t = 1/(C\mu_i)$ where μ_i is the mutation rate in region i. Since the mutation rate is exponentially distributed, with mean μ, we can then explicitly calculate the value of $N_A(0)$ which is the number of regions in sub-genome A which have not mutated yet at time t:

$$\begin{aligned} N_A(0) = P(t\mu_i C < 1)\, R \quad &= P\left(\mu_i < 1/(Ct)\right) R \\ &= \left(1 - e^{-1/(Ct\mu)}\right)\frac{L}{2C} \end{aligned} \tag{5.5}$$

Thus

$$M(r) = \frac{\lambda K L}{\mu}\frac{1}{r^3} + \frac{L^2}{C}\frac{3}{r^4}\left(1 - e^{-1/(Ct\mu)}\right)^2, \tag{5.6}$$

so the term in r^{-4} dominates if

$$\frac{\lambda K}{\mu} \ll \frac{L}{C}\frac{1}{r}\left(1 - e^{-1/(Ct\mu)}\right)^2. \tag{5.7}$$

From the last equation, we obtain that at any time t, the $\alpha = -4$ power-law regime holds for length up to $r_c(t)$, with

$$r_c(t) = \frac{\mu L}{\lambda K C}\left(1 - \exp[-1/(Ct\mu)]\right)^2. \tag{5.8}$$

Similarly, if the $\alpha = -4$ power-law behavior holds up to a length r, we can calculate the time t_c elapsed since the WGD event as:

$$t_c = -\left\{C\mu \log\left[1 - \left(\frac{r\lambda K C}{\mu_c L}\right)^{1/2}\right]\right\}^{-1}. \tag{5.9}$$

Surprisingly, the value of t_c decreases with the mean size C of a region. This is due to the fact that the most important parameter here is the number of regions. As the size of regions increases, the number R of regions decreases (if the length of the genome stays the same).

5.2.1 Simulations

Using a model similar to the one described in Section 4.4.2, we simulated sequences undergoing a WGD event. We started from an iid sequence of length $L/2$. We then let the sequence evolve for a time t_0 according to the model of Section 3.2.2 to generate genome I. We then duplicated this genome to produce genome A and B, and simulated the divergence between the two sub-genomes as if they were independent genomes, according to model of Section 4.4.2 for a time t_1. At the end of the simulation, we concatenated the two sub-genomes and computed the MLD from the self-alignment of the full genome. We show the results of our simulations for different time t_1 of evolution on Fig. 5.2

The results obtained on sequences simulated this way agree well with our calculation for the time of transition between the two regimes. Notably, one can see on fourth panel of Fig. 5.2 that the transition for a length $r_c = 200$ occurs at time $t_c = 1$.

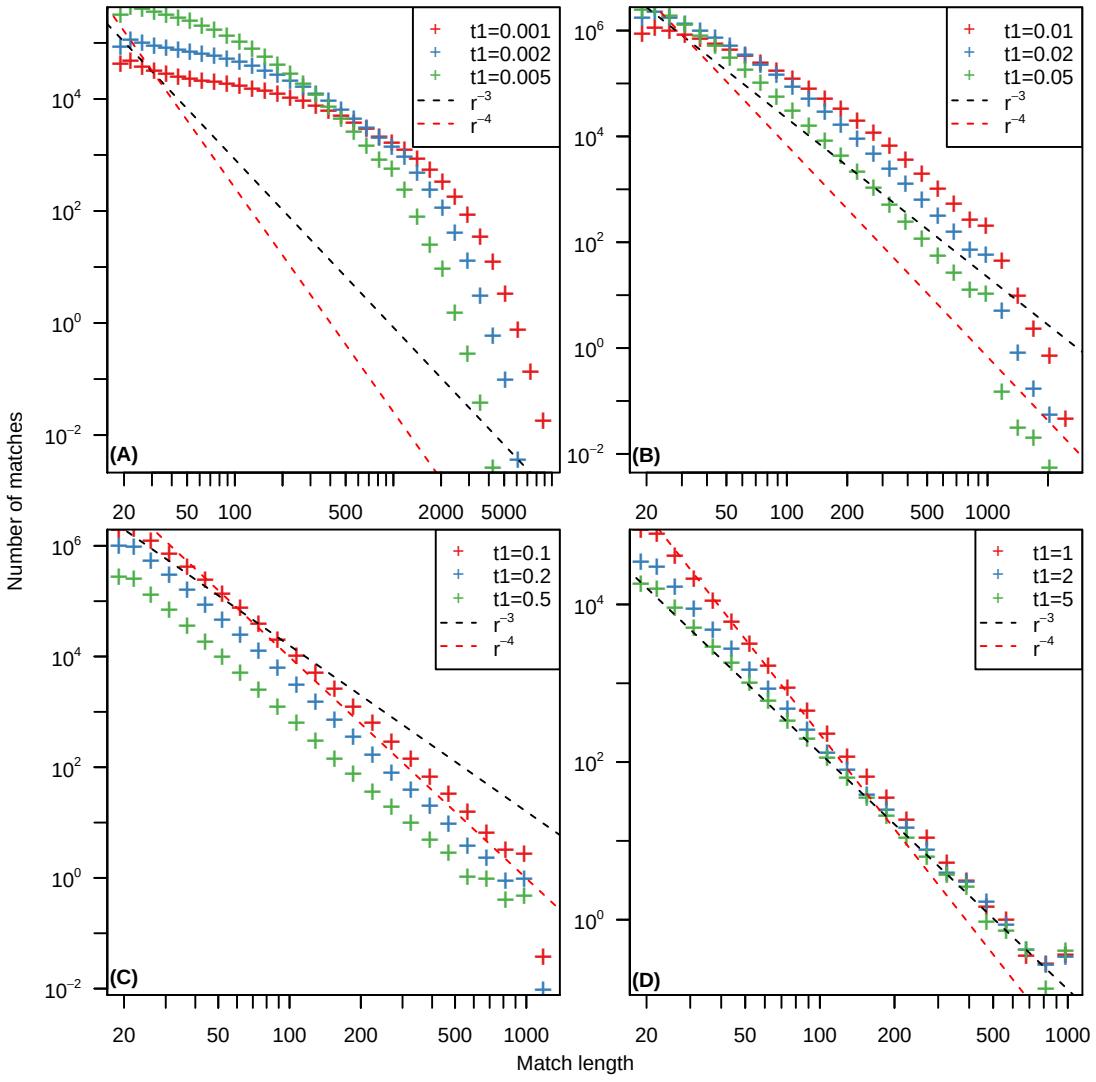

FIGURE 5.2: The MLD computed for the self-alignment of sequences simulated with the model described in Section 5.2.1 for different evolution times t_1. For all simulations, parameters are $\lambda = 10^{-5}$, $\mu = 1$, $K = 1000$, $C = 1000$ and initiation time $t_0 = 1$. The dashed lines represent power-laws with exponent $\alpha = -3$ and $\alpha = -4$. All the empirical data are represented using logarithmic binning.

5.3 Discussion and Limitations

In this section, we have shown that after a WGD, the MLD computed from the self-alignment of a genome can result in an $\alpha = -4$ power-law. Interestingly, we have shown on Fig. 3.4 (on page 71) that the MLD computed from the self-alignment of several species did result in an $\alpha = -4$ power-law. Indeed, for two of these "-4"-species (*Arabidopsis thaliana* and Zebrafish), many evidences for a recent WGD have been described [125, 126].

Note that for the estimation of the transition point between the two regimes, we used several strong hypotheses. Especially, as stated in the last section, we cannot estimate the value of C (we can only estimate a lower bound for C). More importantly, we need an hypothesis on the distribution of mutation rates over different regions to try to estimate the value of $N_A(0)$. From our analysis, we cannot estimate these parameters for real genomes.

If we however apply our formula to try to estimate t_c using Eq. (5.9) in the human genome, using common estimates for the mean mutation rate ($\mu = 1.5\,10^{-9}$ per bp per year) and for the mean length of a segmental duplication ($K = 10^4$ bps), our estimate from Chapter 3 for the segmental duplication rate ($\lambda = 1.5\,10^{-13}$), and the lower bound for C from Section 4.4.3 ($C > 300$ bps), we can estimate an upper bound for the value of $t_c < 4\,10^8$ years for matches of length $r = 100$ bps. According to various studies reviewed in Kasahara [127], 2 rounds of WGD occurred between 450 and 550 million years ago, in the early history of vertebrates, in good agreement with our estimate.

However, WGD events are known to strongly impact the evolution of genomes. Notably, they are often associated to speciation events [58, 128]. Thus, they are often associated with complex functional evolutionary processes that might not be well described by our model. For instance, after a WGD, many duplicated genes are lost in a process known as fractionation, and it has been found that one of these loss occurred preferentially in one of the two subgenomes, see Thomas, Pedersen, and Freeling [59], Woodhouse, Schnable, Pedersen, Lyons, Lisch, Subramaniam,

and Freeling [129]. More generally, it has been reported that WGD events were followed by extensive chromosomal rearrangements [130–133], and that the two sub-genomes might not evolve independently [134, 135]. In our model, we do not take into account all these complex behaviors.

Chapter 6

Comparison of Coding Sequences

In this Chapter, we study only the coding part of genomes, namely, the exomes. We show that the MLDs computed from the comparisons of the exomes of two distinct species exhibit a different behavior as the one computed from the comparative alignment of full genomes, that is, an $\alpha = -5$ power-law. We then discuss possible models that might explain these observations, and that should be further investigated.

6.1 Comparing the Exome of Different Species

While comparing the RepeatMasked genomes of several species, we observed that MLDs were shaped as $\alpha = -4$ power-laws for a wide range of comparisons (see Chapter 4). Interestingly, when we compared the non-coding part of any two genomes, we obtained highly similar distributions, and we found that the majority of matches obtained when comparing genomes were part of the non-coding regions of both genomes. We then computed the comparative alignment of the coding part of both genomes (namely, we concatenated all exons of each genome in one long sequence, the exome, and computed the MLD from the comparison of these two sequences). To avoid creating irrelevant matches, exons were separated by "N"s in this sequence.

Fascinatingly, the MLD computed from this comparison results again in a power-law distribution, but this time, the exponent of the power-law is $\alpha = -5$. As in the case of comparative alignment discussed in Chapter 4, the distance between the two species has to be large enough so that the distribution is observed. The closest pair for which we found the $\alpha = -5$ power-law was the Human/Mouse pair, while the two most distantly related pairs were Mouse and Chicken (see Fig. 6.1 and Fig. 6.2). As in the case of whole genome comparison, this distribution is not observed in the comparisons of species which are too close to each other (namely, for the pairs Human/Chimp, Mouse/Rats but also, Human/Dog). For these pairs, the number of matches decreases exponentially with the length of matches. Compared to the distribution described in Chapter 4, the exponential behavior is observed for a broader range of pairs. For instance, the MLD of the Mouse/Rat full genome comparison exhibits a power-law tail while its exome counterpart is exponential, and similarly for the Human/Dog pair. Like in the case discussed in Chapter 4, the power-law behavior is lost when the two species are too distantly related (see Fig. 6.1 and Fig. 6.2), and the distance from which no power-law can be observed seems roughly the same for both cases. To sum up, as for full genome comparisons, we observe 3 regimes for the exome comparisons, but the range of distances where the $\alpha = -5$ power-laws holds is more narrow.

Of course, it is not totally unexpected to observe different results for the comparison of the exome and the comparison of non-coding sequences. Notably, due to the degeneracy of the genetic code, the mutation rate in exons is known to be much higher (roughly 10 times) at the 3rd position than at 1st and 2nd positions of a high number of codons. The first consequence of this property of exons is that we observe much more exact matches of length $r = 3n + 2$ than matches of length $r = 3n + 1$ and $r = 3n$ (with $n \in \mathbb{N}$), such that the distribution show a periodic pattern, with periodicity 3. One cannot observe this trend on data represented with the logarithmic binning because bins overlap several lengths. The periodicity can be observed on Fig. 6.3 representing the Human/Mouse exome comparison without the logarithmic binning.

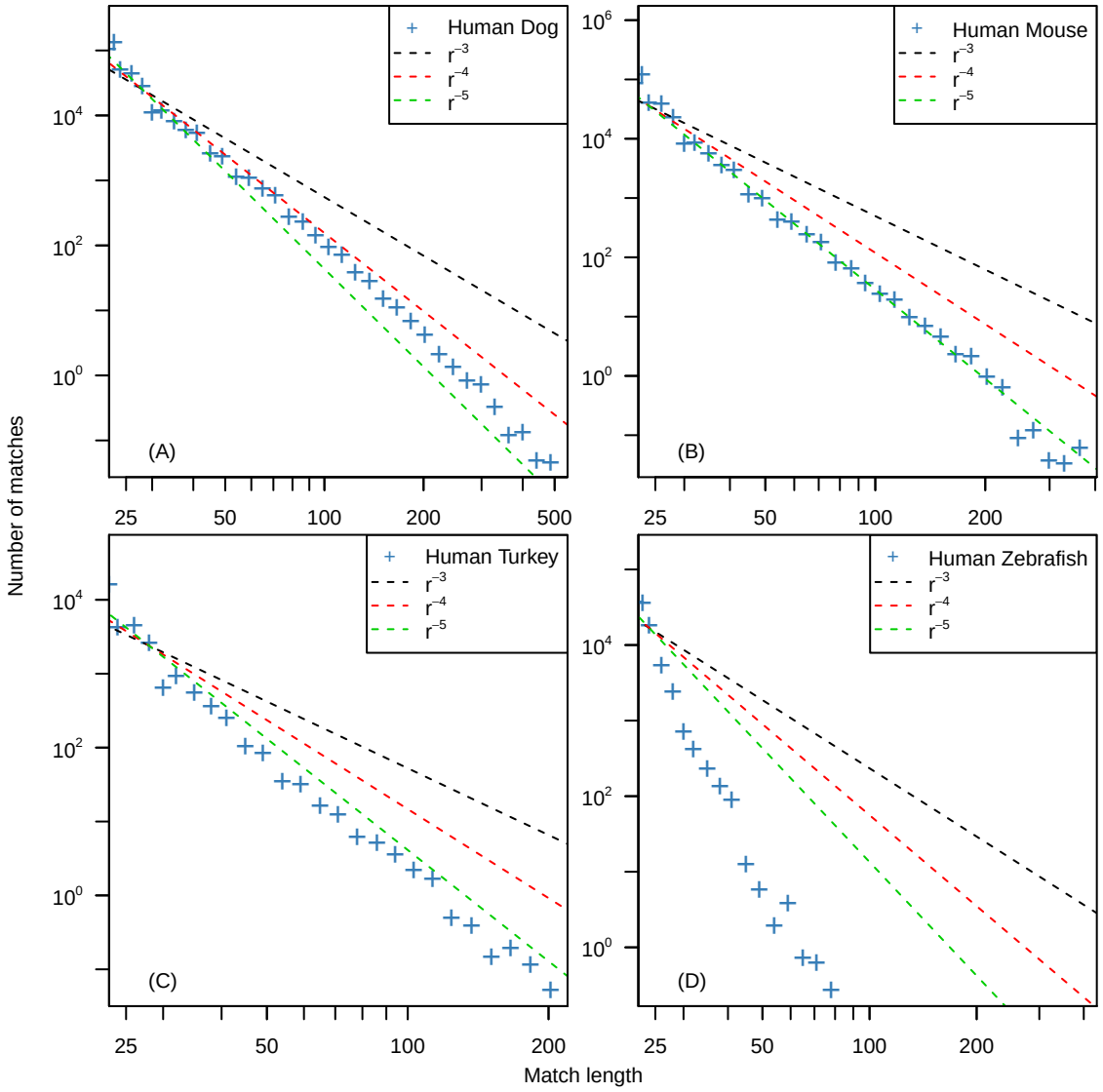

FIGURE 6.1: MLD computed from the comparison of the human exome of with the exome of several species. The four species are chosen to represent the three different states described in the main text. In all four panels, dashed lines represent power-law distribution with exponent $\alpha = -3$, $\alpha = -4$ and $\alpha = -5$, and empirical data are represented using logarithmic binning. MLDs represented are computed from the comparison of the exomes of (A) Human and Dog (B) Human and Mouse (C) Human and Turkey (D) Human and Zebrafish

To understand the impact on the MLD of the fact that the mutation rate varies at the bp level, we generated 3 sets of sequences from the exome of both Human and Mouse, that were defined as follow: we first downloaded from the `Ensembl` database only the transcripts that were known to be translated into a protein. Using these transcripts, we concatenated all first, second and third base pairs separately in 3 different chimeric sequences (designed by SeqPos1, SeqPos2 and

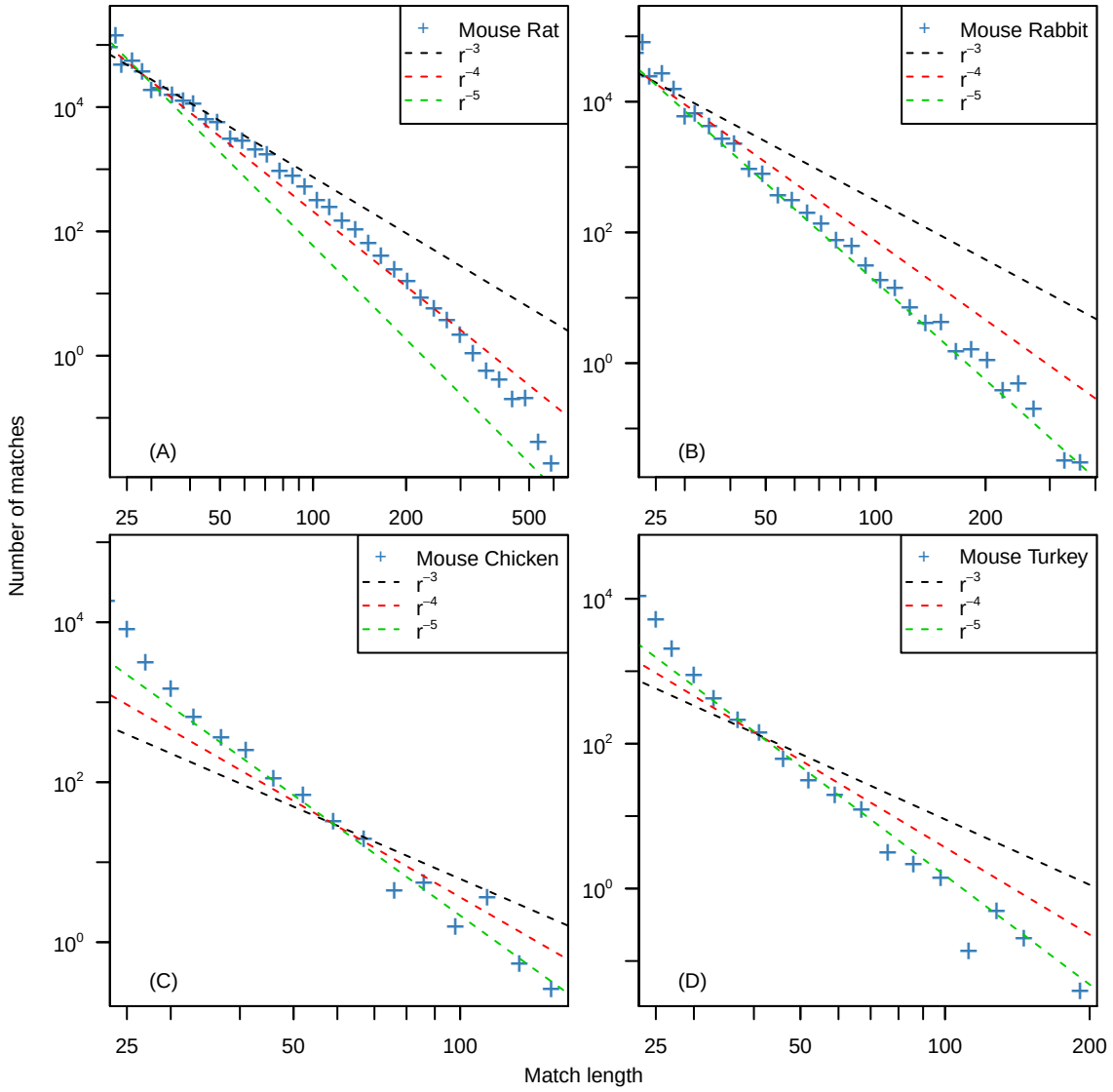

FIGURE 6.2: MLD computed from the comparison of the exome of several species. In all four panels, dashed lines represent power-law distribution with exponent $\alpha = -3$, $\alpha = -4$ and $\alpha = -5$, and empirical data are represented using logarithmic binning. MLDs represented are computed from the comparison of the exomes of (A) Mouse and Rat (B) Mouse and Rabbit (C) Mouse and Chicken (D) Mouse and Turkey.

SeqPos3 respectively in the following). This way, we obtained 3 sequences where the variation of the mutation rate at the base pair level has been eliminated. From these sequences, we computed the MLD from the comparison of each sequence to its homologous in the other species. These 3 MLDs are shown on panels (A,B and C) of Fig. 6.4. Surprisingly, the comparison of the third positions of exomes exhibits an $\alpha = -5$ power-law, while in the two other comparisons, the number

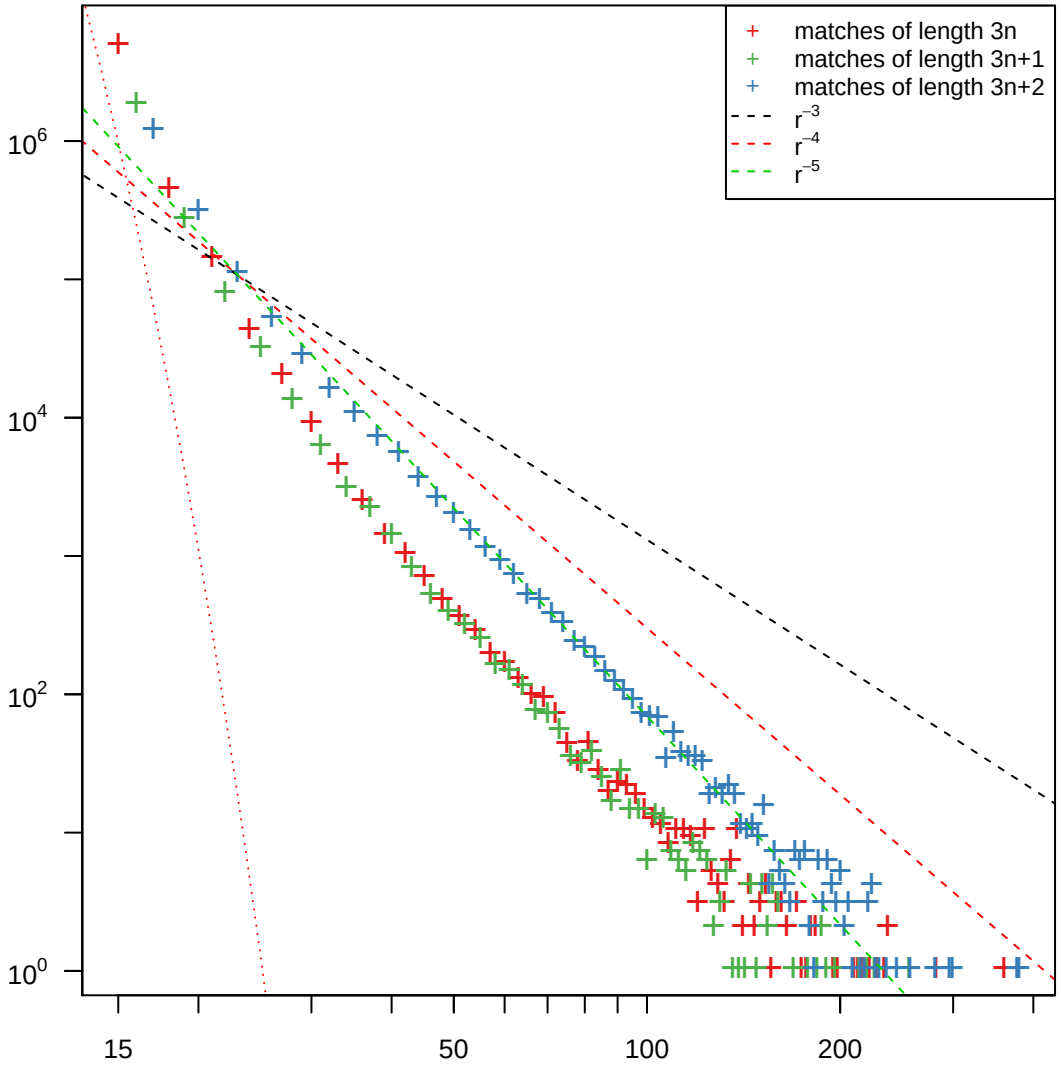

FIGURE 6.3: MLD computed from the comparison of human and mouse exome, represented without logarithmic binning. 3 different colors are used to represent matches of length $3n$, $3n+1$ and $3n+2$. In all four panels, dashed lines represent power-law distribution with exponent $\alpha = -3$, $\alpha = -4$ and $\alpha = -5$.

of matches decay exponentially with the length of the matches. As expected, the dominant divergence process between the exomes is the mutation occurring at the third position of exons. From this experiment, we can conclude that the only relevant mutational process in the case of exome comparison are mutations occurring in the third base pairs of codons.

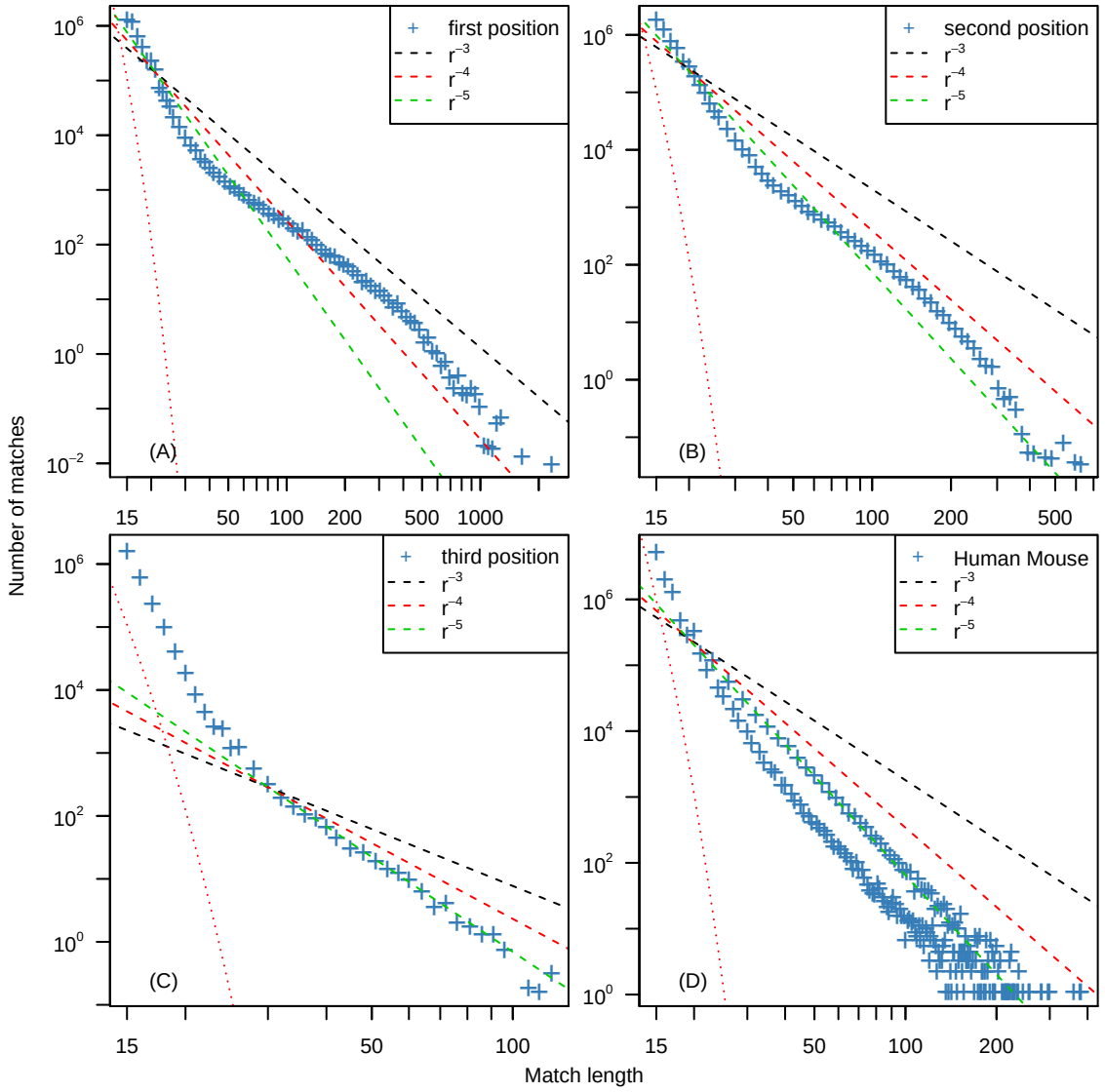

FIGURE 6.4: MLD computed from the comparison of :(A) SeqPos1 of Human and SeqPos1 of Mouse, (B) SeqPos2 of Human and SeqPos2 of Mouse (C) SeqPos3 of Human and SeqPos3 of Mouse. These 3 MLDs are represented using a logarithmic binning. (D) The MLD computed from the comparison of the full Human and Mouse exome, represented without logarithmic binning. In all four panels, dashed lines represent power-law distributions with exponent $\alpha = -3$, $\alpha = -4$ and $\alpha = -5$.

6.2 Theoretical Application of the Divergence Model

In this subsection, we apply the model developed in Section 4.4.1 to the case of exome comparison. In theory, the fact that we now consider only exons, and not

the full genome should not impact our calculations of $N(\tau)$, such that Eq. (4.7) still holds:

$$N(\tau) = \int_0^\tau N_A(\tau - \tau_B) N_B(\tau_B) d\tau_B. \tag{6.1}$$

As we are interested in long matches, and as long matches mostly stem from the comparison of sequences which exhibit a small divergence, we compute the Taylor expansion of $N(\tau)$ next to 0. Assuming that the third derivative of $N(\tau)$ and the Taylor expansion exists, we find:

$$\begin{aligned} N(\tau) &= N(0) + \left.\frac{dN(\tau)}{d\tau}\right|_{\tau=0} \tau + \left.\frac{d^2N(\tau)}{d\tau^2}\right|_{\tau=0} \frac{\tau^2}{2!} \\ &\quad + \left.\frac{d^3N(\tau)}{d\tau^3}\right|_{\tau=0} \frac{\tau^3}{3!} + \mathcal{O}(\tau^4). \end{aligned} \tag{6.2}$$

Using Leibniz formula to calculate the first three derivatives of $N(\tau)$ from Eq. (6.1), it follows that:

$$\begin{aligned} N(\tau) &= 0 + N_A(0)N_B(0)\tau + [N'_A(0)N_B(0) + N_A(0)N'_B(0)] \frac{\tau^2}{2!} \\ &\quad + [N''_A(0)N_B(0) + N'_A(0)N'_B(0) + N_A(0)N''_B(0)] \frac{\tau^3}{3!} + \mathcal{O}(\tau^4). \end{aligned} \tag{6.3}$$

In this case again, we can apply Eqs. (3.2) and (3.6), and replacing the mean length of a duplication K by the mean length of a conserved region C. Note that in that case, the value of C is constrained by the length of exons (see Section 6.5 for a discussion on this parameter). Then, integrating by parts, it follows that the MLD is given by:

$$\begin{aligned} M(r) &= N_A(0)N_B(0)\frac{6(C-r+2)}{r^4} + [N'_A(0)N_B(0) + N_A(0)N'_B(0)] \frac{12(C-r+2)}{r^5} \\ &\quad + [N''_A(0)N_B(0) + N'_A(0)N'_B(0) + N_A(0)N''_B(0)] \frac{20(C-r+2)}{r^6} + \mathcal{O}(r^{-7}). \end{aligned} \tag{6.4}$$

Depending on which term dominates the sum in Eq. (6.4), one can predict which power-law tail will be observed in the MLD of the comparative alignment.

In the case of comparative alignment of whole genomes one clearly observes the

power-law with an exponent $\alpha = -4$. This indicates that the term $N_A(0)N_B(0)$ is dominant. This is expected because of the existence of ultra-conserved elements in genomes, such that the distribution of mutation rate does not vanish at zero, and thus $N_A(0)N_B(0) > 0$.

In exons however, due to the redundancy of the genetic code, the mutation rate is much higher at the third position of exons than in the rest of the sequence. Thus, we do not expect to observe non-mutating sequences, i.e. $N_A(0) \simeq 0$ and $N_B(0) \simeq 0$. In that case, the term in r^{-4} of Eq. (6.4) vanishes. But in that case, the term in r^{-5} also vanishes, and we switch directly to an $\alpha = -6$ power-law. Thus, according to our theoretical calculations, we should never observe an $\alpha = -5$ power-law.

Still, it might be that constraints on the exons are strong enough so that $N_A(0) > 0$ or $N_B(0) > 0$, but that we still expect an $\alpha = -5$ power-law. In that case, we can calculate conditions that would result in $N(\tau) \sim \tau^2$, and that would imply $M(r) \sim r^{-5}$.

6.3 Theoretical Calculation of the Value of $N(\tau)$

Symmetrical Mutation Rate Distributions — Let us first assume that the mutation rate distribution is the same in the exome of both species (i.e. $N_A = N_B$). In this case, Eq. (6.4) becomes:

$$\begin{aligned} M(r) &= N_A(0)^2 \frac{6(C-r+2)}{r^4} + 2\left[N'_A(0)N_A(0)\right] \frac{12(C-r+2)}{r^5} \\ &\quad + \left[2N''_A(0)N_A(0) + N'_A(0)^2\right] \frac{20(C-r+2)}{r^6} + \mathcal{O}(r^{-7}). \end{aligned} \tag{6.5}$$

We have already seen that if the term in r^{-4} vanishes, then the term in r^{-5} also vanishes and we expect an $\alpha = -6$ power -law.

If $N_A(0) \neq 0$, we could obtain an $\alpha = -5$ power-law behavior if the term in r^{-5} dominates the term in r^{-4} and the term r^{-6} , namely if:

$$\begin{cases} (i) & 2N_A(0)N'_A(0)\, r^{-1} \gg N_A(0)^2 \\ (ii) & 2N_A(0)N'_A(0) \gg [2N''_A(0)N_A(0) + N'_A(0)^2]\, r^{-1}. \end{cases} \tag{6.6}$$

Condition (i) is equivalent to $N_A(0) \ll 2N'_A(0)r^{-1}$ for all values where the $\alpha = -5$ power-law holds, i.e. for $r \in [20, 200]$. If $N''_A(0) > 0$ then it follows from condition (ii) that $N_A(0) \gg N'_A(0)r^{-1}$ for $r \in [20, 200]$ and the two conditions are contradictory. Thus, again, we expect the exponent of the power-law to be either equal to -4 or to -6.

If $N''_A(0) < 0$, condition (ii) results in $N''_A(0) \ll rN'_A(0)\left[1 - \frac{N'_A(0)}{2N_A(0)r}\right]$, and using condition (i) we get:

$$N''_A(0) \ll -N'_A(0)\frac{N'_A(0)}{2N_A(0)} \ll -rN'_A(0),\ \forall\ r \in [20, 200] \tag{6.7}$$

which implies strong constraints on the distribution of N_A in the neighborhood of $\tau = 0$. Recall that $N_A(\tau)$ represents the number of sequences in species A which are at an evolutionary distance τ from their common ancestor, which is equivalent to the distribution of mutation rates accross the different exons of sequence A. The condition (6.7) implies strong constraints on the mutation rate. Although we cannot rule out such constraints, there is no simple biological reason that would justify them, especially as they should apply on the mutation rate in exons, and not in the non-coding part of the genome.

Asymmetrical Mutation Rates Distribution — However, the $\alpha = -5$ power-law could also result from the asymmetry of mutation rates. In this case, we would

have that $N_A \neq N_B$, and thus condition Eq. (6.6) turns into:

$$\begin{cases} (iii)\,[N_B(0)N'_A(0) + N_A(0)N'_B(0)]\,r^{-1} \gg N_A(0)N_B(0) \\ (iv) N_A(0)N'_B(0) + N'_A(0)N_B(0) \gg \dfrac{N''_A(0)N_B(0) + N'_A(0)N'_B(0) + N''_B(0)N_A(0)}{r} \end{cases} \tag{6.8}$$

If we now consider a simple scenario where $N_B(0) = 0$ and $N_A(0) \neq 0$, condition (iii) is trivial and condition (iv) results in: $r > \frac{N'_A(0)}{N_A(0)} + \frac{N''_B(0)}{N'_B(0)}$. Again, there is no obvious reason that could lead to such properties of mutation rate distributions, but one cannot rule it out.

However, assuming an asymmetry between species leads to a contradiction. It is easy to see it in the case where $N_A(0) = 0$ while $N_B(0) > 0$. Let us introduce a third species, species C. Then, either $N_C(0) = 0$ or $N_C(0) > 0$. In the first case, only the comparison of B and C is supposed to result in an $\alpha = -5$ power-law while the comparison of A to C would be an $\alpha = -6$ power-law. In the second case however, only the comparison of C to A would follow an $\alpha = -5$ power-law, while the comparison of B to C would be an $\alpha = -4$ power-law. Hence, in both cases, we expect the comparisons A/C and B/C to give different results. This contradicts empirical data. Indeed we found many triplets of species where all one-to-one exome comparisons exhibit an $\alpha = -5$ power-law (see for instance all comparisons from the species Human, Mouse, Chicken and Rabbit on Figs. 6.1 and 6.2). Thus, the $\alpha = -5$ power-law does not result from an asymmetry in the distribution of mutation rates between species.

6.4 Investigating Different Exon Subclasses

As the simple models developed did not succeed at explaining the observations obtained from the comparison of the exome of different species, we investigated the contributions of the different classes of exons on the shape of the MLD.

In eukaryotic genomes, some genes are composed of a single exon (single exon genes) while some are composed of several exons (also called multi-exon genes). In the latter case, the entire gene is transcribed, and the primary mRNA molecule is then spliced to remove introns. When this is the case, depending on various physiological conditions of the cell, some exons can also be spliced. This phenomenon, know as alternative splicing, is a way for the organism to produce different proteins, depending on the function of the cell.

As a consequence, the constraints that shape the evolution of an exon might be different if it belongs to a single exon gene, or to a multiple exon gene. In the latter case, whether an exon is the first, the last or located in the middle of a gene might also affect its evolution.

In the following, for simplicity, we focus only on the Human/Mouse exome comparison. To see whether these structural properties have an impact on our analyses, we constructed different sets of exons:

– Set 1: Exons which come from Single exon genes,

– Set 2a: Exons which come from multi-exons genes and which are the first exon of the gene,

– Set 2b: Exons which come from multi-exons genes and which are the last exon of the gene,

– Set 2c: Exons which come from multi-exons genes and which are neither the first nor the last exon of the gene (we also refer to them as middle exons).

We build each different sets in both species, and then concatenated all the exons of each set (separating them with an "N" letter) to obtain 4 different sequences for each species. We then computed the MLDs from the comparison of the same set from different species ("parallel comparison"), Fig. 6.5 and of different sets from different species ("cross comparisons"), see Fig. 6.6.

Surprisingly, we observed that these different comparisons did not all result in an $\alpha = -5$ power-law. Indeed, for parallel comparisons, only the comparison of

	Human		Mouse	
	Length	Percentage	Length	Percentage
Single Exons	10.6 Mbps	10.3%	6.7 Mbps	8.2%
Middle Exons	65.4 Mbps	63.8%	48.3 Mbps	59.2%
First Exons	19.7 Mbps	19.2%	17.4 Mbps	21.3%
Last Exons	6.8 Mbps	6.7%	9.2 Mbps	11.2%
Total	102.6 Mbps	100%	81.6 Mbps	100%

TABLE 6.1: This table summarize the length and proportion of the full exome of sequences constructed on all four subsets in each species

middle-exons exhibits this behavior. Interestingly, this set is by far the largest in terms of number of base pairs in both species (see Table 6.1), and the parallel comparison of this set in both species is responsible of most of the matches. These observations raise the question of whether, a peculiar feature of middle-exon genes is responsible for the behavior we observe. On the other hand, the MLDs obtained from the first and last exons seem to exhibit an $\alpha = -4$ power-law, in agreement with the model developed about comparative MLD of non-coding DNA, while the MLD computed from the comparison of single exons seems to exhibit an even higher exponent. Note however that while extracting subsets of the exome, we dramatically (especially for sets 1, $2a$ and $2b$) reduce the size of the two sequences to compare, and, as a consequence, the number of matches. Thus, there might not be enough matches left to observe clean power-law distributions, which could explain the apparent strange behaviors we obtain.

Another important feature of exons that might affect mutation rates is the number of copies in which each exon is present. Indeed, several processes specific to duplicated genes are known to affect their evolutionary fate (such as gene conversion or subfunctionalization [42, 136]).

To classify exons according to this property, we repeated the procedure of Section 4.3 on Human and Mouse exomes: using the self-alignments of both Human and Mouse exomes, we separated the exons in two sets. In the "paralog" set, we

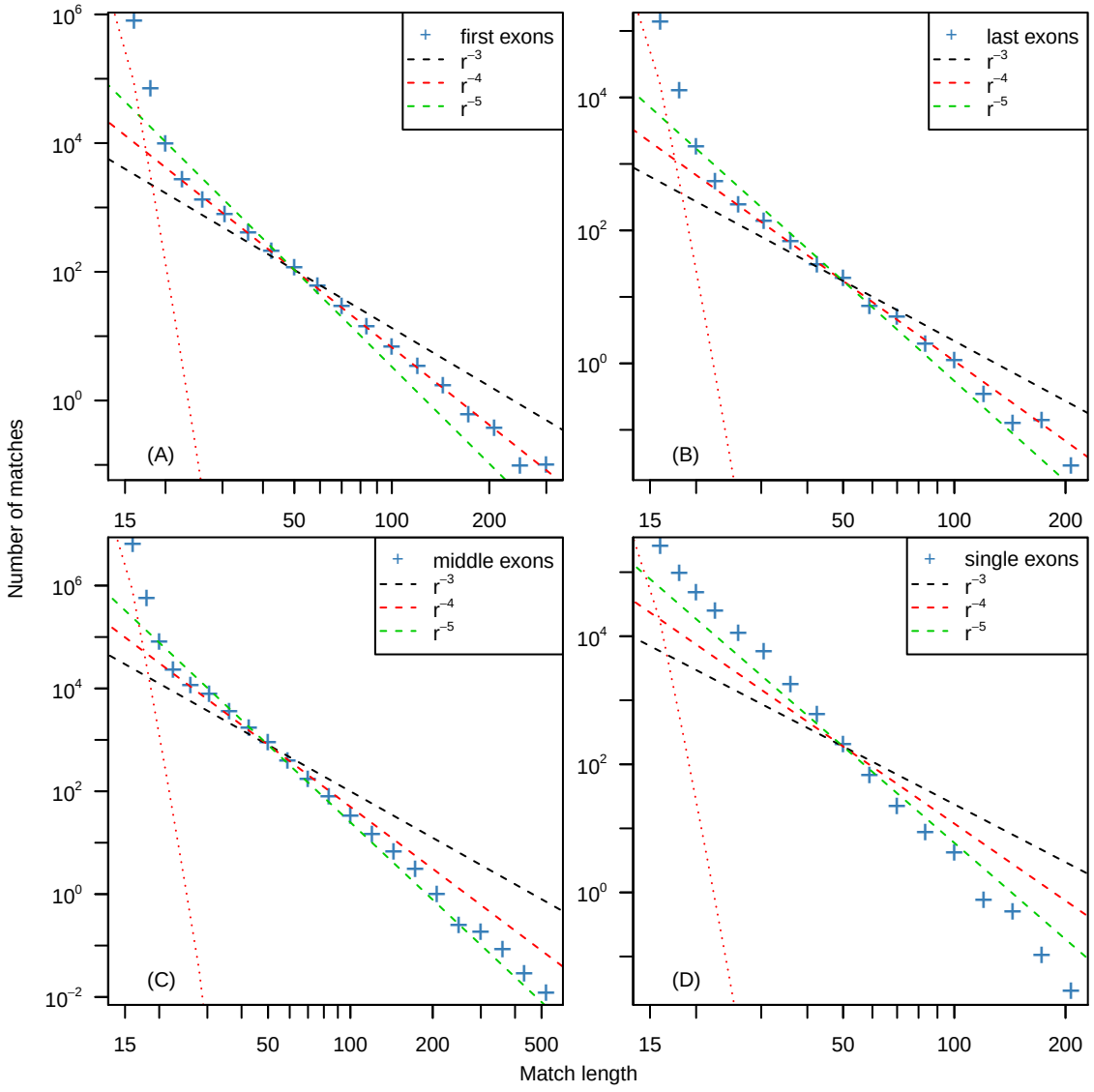

FIGURE 6.5: MLD computed from the parallel comparisons of the different subsets constructed on the Human and Mouse Exomes.

only kept exons that match (matches are defined as exact match of length $r \geq 20$) with another exon in the same genome. All other exons are unique, and we put them in the "unique set". For both species, the size of each subset is roughly half the size of the full set (the unique sets are of length 50 Mbps and 55 Mbps while the paralog sets are of length 32 Mbps and 48 Mbps for Mouse and Human respectively).

We then again computed the MLDs from the comparison of sequences produced from the concatenations of each set. Here again, we compute both "parallel"

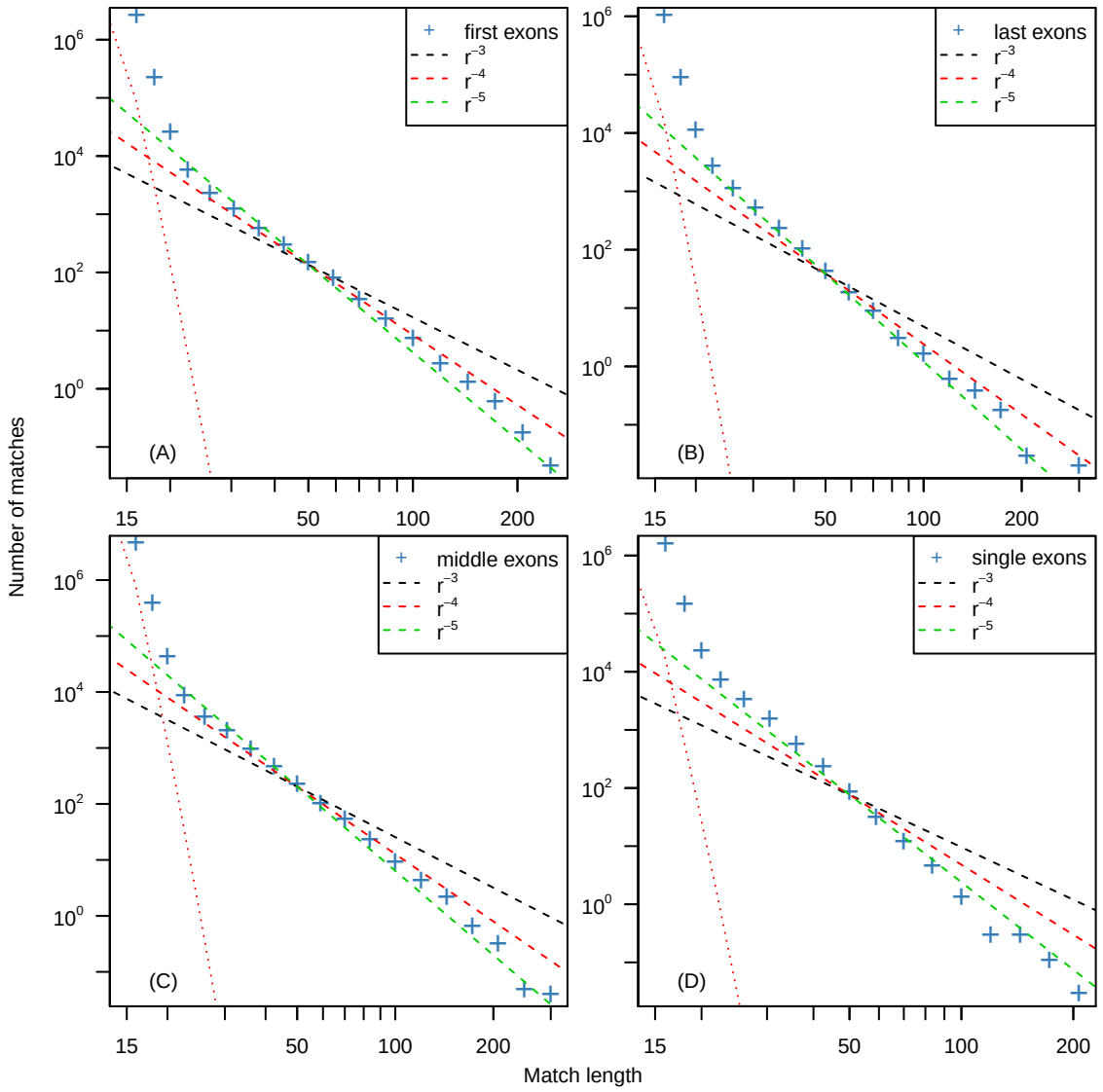

FIGURE 6.6: MLD computed from the cross comparisons of the different subsets constructed on the human and mouse Exomes. (A) Human first exons against all the other sets of exons in Mouse (B) Human last exons against all the other sets of exons in Mouse (C) Human middle Exons against all the other sets in Mouse (D) Human last exons against all the other sets of exons in Mouse.

and "crossed" comparison, and show the resulting MLDs on Fig. 6.7. In terms of matches, the largest contribution comes from the comparison of the two non-unique sets (397000 matches) while the comparison of the two unique sets results in 148000 matches and the sum of two cross comparisons in 143000 in total.

In all comparisons, the MLD seems to exhibit a power-law with exponent $\alpha = -5$ although, as in the previous case, the small size of the sets might not allow to draw clear conclusion about the distribution we observe, and the comparison that results

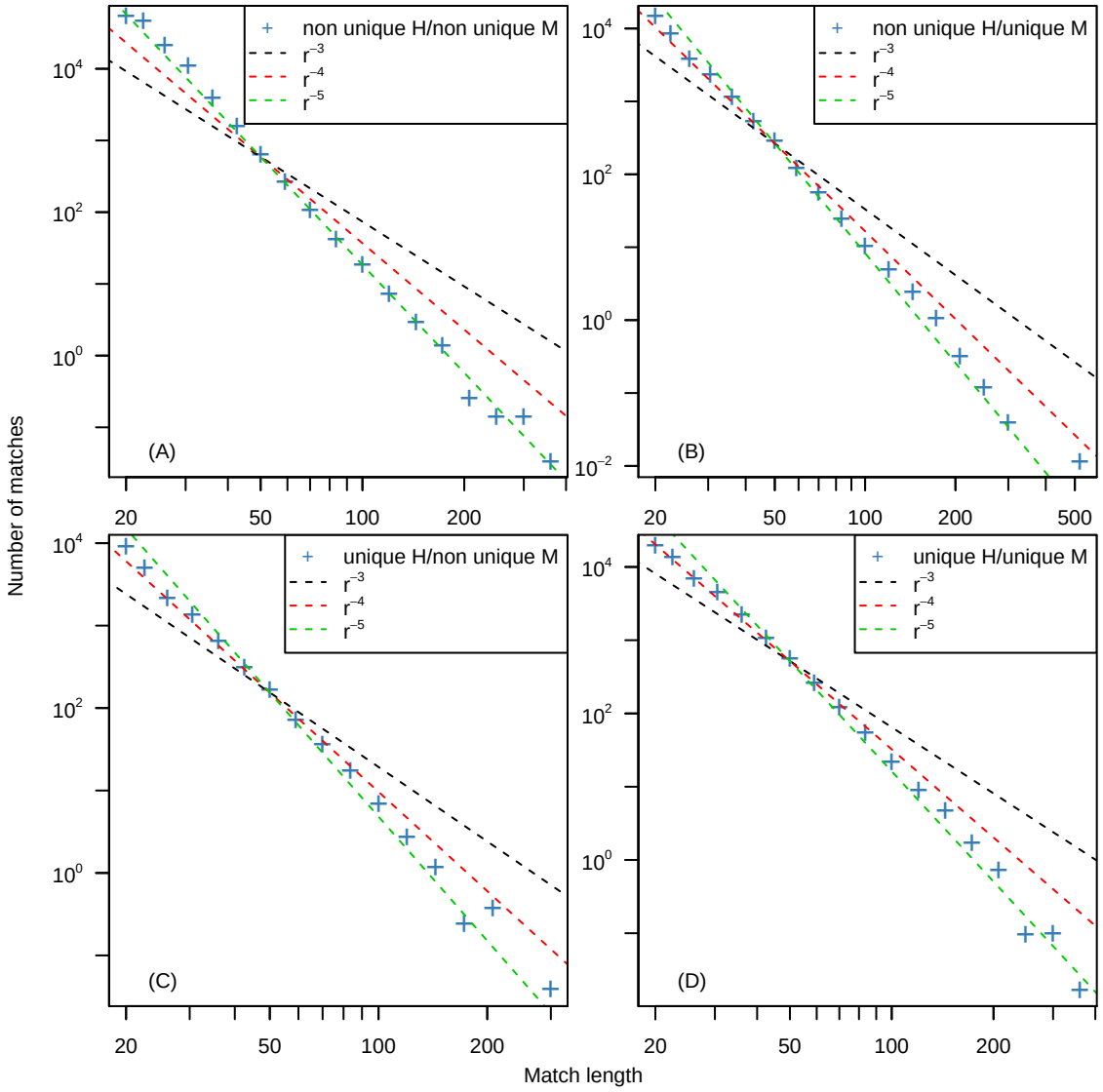

FIGURE 6.7: MLD computed from the parallel and cross comparisons of the unique set and the paralog set of human and mouse exomes.

in the highest number of matches is also the one where the $\alpha = -5$ power-law behavior is the clearest.

6.5 Conclusions

From all the observations that have been reported in this Chapter, we still cannot explain why the comparison of exomes results in an $\alpha = -5$ power-law distribution.

Nevertheless, given our observations, we can point three ingredients that could explain our observations.

Hypotheses Regarding $N(\tau)$ — The first set of hypothesis involves the value of $N(\tau)$. Assuming that (i) due to the high mutation rate on the third base pair of codons, the distribution of mutation rate vanishes in zero in both species, and that (ii) the mutation rate distribution is the same in both species, we reached the conclusion that we expect the MLD to be distributed as an $\alpha = -6$ power-law. We have seen that we could explain the observation of an $\alpha = -5$ power-law distribution by relaxing one or both of these hypotheses. However, both (i) and (ii) seem to be natural and biologically valid hypotheses, that we have not been able to disprove up to now. We were even able to show that supposing that the $\alpha = -5$ MLD power-law was explained by the asymmetry of mutation rates distribution between the species could not explain all our observations.

Hypotheses Regarding $m(r, \tau)$ — In this thesis, we assumed that mutations were, on average, distributed uniformly and independently along sequences. As a consequence, the number of matches expected between 2 sequences follows the broken stick model expectation given by Eq. (3.2). If point mutations in exons are not (on average), uniformly distributed (if, for instance, mutations do not occur independently from each others), it could be that the stick breaking model does not apply. Let us assume that the distribution of the number of matches of length r for two sequences exhibiting a divergence τ is equal to $m_e(r, \tau)$, and that

$$m_e(r, \tau) \neq (2\tau + (C - r)\tau^2) \exp(-r\tau). \tag{6.9}$$

In this case, Eq. (3.6) still applies, such that we have:

$$M(r) = \int_0^\infty m_e(r, \tau)\, N(\tau)\, d\tau. \tag{6.10}$$

In that case, several models could lead to an $\alpha = -5$ power-law. Supposing that $N(\tau) \sim \tau^3$, as expected for the comparison of the exome of diverging species, we

would have

$$M\left(r\right)=\int_{0}^{\infty}m_{e}\left(r,\tau\right)\tau^{3}d\tau. \tag{6.11}$$

Now if for instance

$$m_{e}(r,\tau)=\frac{m(r,\tau)}{\tau}=\frac{(2\tau+(C-r)\tau^{2})\exp(-r\tau)}{\tau}, \tag{6.12}$$

then we would have

$$\begin{aligned} M(r)=\int_{0}^{\infty}N(\tau)m_{e}(r,\tau)\,d\tau &= \int_{0}^{\infty}\tau^{2}m(r,\tau)d\tau \\ &\sim r^{-5}, \end{aligned} \tag{6.13}$$

according to our observations. However, many different values of $m_e(r,\tau)$ could lead to the same result, and one would need to justify these distributions. Note that we tried to measure an empirical value of the distribution of $m_e(r,\tau)$ and the results we have obtained were always in agreement with the distribution expected under the stick breaking model. Moreover, we have seen in Section 6.1 that the MLD computed from the comparison of third positions of the codons, which are supposed to evolve more or less neutrally, still exhibited the same $\alpha=-5$ power-law. While one could imagine that mutations that affect sequences under strong evolutionary constraints are not uniformly and independently distributed, the non-uniformity of mutations that occur at the third base pair of codons seems much more unlikely.

Hypotheses Regarding the Exon Length Distribution — The last parameter we have not discussed yet, and that could impact the shape of the distribution is the length distribution of exons themselves. While in the case of full genome comparisons, C was representing the length of a region of constant mutation rate, in the case of exome comparison, C is constrained by the length of the exons. In the present case, we assumed that all exons are of the same length C. We have seen that this simplistic modeling has no consequence on the expected shape of the MLD since the relationship between $M(.)$ and C is always purely linear. However,

we always assumed that the length of matches was smaller than the value of C. As many exons are small (and especially, smaller than the length of the longest match), the value of C might play a more important role here. For instance, one fifth of exons are longer than 300 bps in Human, and exons longer than 300 bps account for half of the total length of the exome. To test the importance of this parameter, we restricted our set of exons to a smaller range of length in both species (i.e. we kept only exons longer or shorter than a certain threshold in both species), and then computed the MLD from the comparison of these subsets. All the MLDs we computed this way still exhibit the $\alpha = -5$ power-law (data not shown). This would indicate that the distribution of C does not influence the shape of the MLD. Note however that when comparing two sequences, the true value of C is the length of the longest homologous subsequence between the two, which can very well be different from the length of the two compared sequences. Although we can use the length distribution of exons in both species as a proxy for the distribution of C, we cannot completely rule out the hypothesis that the distribution of C has an impact on the MLD we study.

Finally, we observed that, depending on their position in the genes, the distributions obtained from the comparison of exons exhibited different behaviors. This could be indicative of more complex dynamics affecting either the mutation rate distribution in exons, or the interaction between mutations (and thus the value of $m_e(r, \tau)$), and that might explain that we observe a MLD shaped as an $\alpha = -5$ power-law.

Chapter 7

Conclusion

7.1 Summary

In this thesis, we have studied properties of match length distributions computed for a various sets of experiments. We developed simple evolutionary models that could successfully explained the deviations from the expected distributions.

In the case of self-alignments, we showed that these deviations were simply resulting from the neutral evolution of duplicated DNA, and that, depending on the mechanism of duplication which is predominant in a genome, different MLDs were expected. In the case of comparative alignments, we found that the deviations were the result of a totally different mechanism. We showed that in comparative alignments, the distribution of mutation rates in the two compared genomes generates the observed power-law MLDs. Thus, unlike in the self-alignment case, these deviations do not result from the neutral evolution of genomes. Finally, trying to apply our evolutionary framework to the comparison of the coding part of genomes, we observed that the processes taking place in the exome of species are more complex, and that additional ingredients – that we have not been able to identify yet – are necessary to explain these deviations.

Table 7.1 summarizes the different power-law behaviors we observed for the MLD when comparing various genomic sequences. It also provides the corresponding evolutionary models we proposed to explain these behaviors.

Experiment		Observed MLD	Model
Self-Alignment			
Human	not RepeatMasked	Exponential	Bursts of TEs
	RepeatMasked	Power-law $\alpha = -3$	Simple SD model Section 3.2
	Processed Pseudogenome	Power-law $\alpha = -4$	Retroduplication Model Section 3.4
	Exome	Power-law $\alpha = -3$	Simple SD Section 3.2
Rabbit / *Arabidopsis Thaliana*/ Zebrafish		Power-law $\alpha = -4$	Retroduplication*/WGD* Section 3.4/ Chapter 5
Comparative Alignments			
	closely related species	Exponential	Simple Stick Breaking Section 4.4.1
	distantly related species	Power-law $\alpha = -4$	Mutation Rate Distribution Section 4.4.1
	Exome	Power-law $\alpha = -5$	Unknown* Chapter 6

TABLE 7.1: A table that summarize the result we obtained. A * indicates that the link between the model and the observation has not been formally established. SD stands for segmental duplication, TE for transposable element and WGD for whole genome duplication.

7.2 Perspectives

As discussed in the introduction, once a model is built, studying deviations from the model can be very useful to get insight on complex behavior, and then to propose new models that fit better the data. With the analysis done in this thesis, we have reached this objective, as we have developed models that explain the most commonly observed MLDs both in self-alignment and in comparative alignments. We have also detected some puzzling behaviors (power-laws with exponent $\alpha = -4$

in the self-alignment of several genomes, or power-laws with exponent $\alpha = -5$ in the comparative alignments of exomes). The next step would then be to identify the biological processes that give rise to these distributions. As the number of genomes available is still constantly growing, more deviations of this kind are expected to be found. Notably, computing a MLD could be used to spot that the quality of the assembly of a genome is poor, and that some transposable elements present in a genome have not been identified.

Appendix A

Extension to the Discrete Case

In this thesis we considered the continuous version of the stick breaking process. In this formalization the stick length r was continuous and the dynamics was described as an integro-differential equation, see Eq. (3.1) in Section 3.1. An equivalent discrete version of the stick breaking model is also known [137]. The size distribution of pieces of a single broken stick is again given by a geometric distribution. The same model, including a continuous generation of new sticks of size K, can also be considered and analytically solved. The solution yields finite size corrections to the power-law behavior for small lengths.

If $M(r,t)$ denotes the number of matching sequences of length r (now discrete) at time t, then we have

$$\frac{\partial M(r,t)}{\partial t} = -2\mu r M(r,t) + 4\mu \sum_{s>r} M(s,t) + \lambda L \delta(r,K). \tag{A.1}$$

The first term on the right hand side expresses the loss of segments due to mutations ($\propto \mu$) in any position in one of the two matching segments ($\propto 2r$). Mutations in matches may also lead to a gain of matches of length r, if a match of longer length $s > r$ is mutated ($\propto \mu$) at position $s+1$ or $r-s$ in one of its two copies ($\propto 4$). This is represented by the second term in Eq. (A.1). In this formulation we disregard the effect that a mutation of one nucleotide in a match of length s leaves behind two matches, which in total have length $s-1$. The third term represents

the gain of matches with length K and with rate λ anywhere along the sequence; $\delta(r, K)$ is the Kronecker delta (i.e. the function which is equal to zero everywhere except for $r = K$, where it takes the value 1).

We are interested in the stationary properties of this model and solve the differential Eq. (A.1) for the stationary state, i.e. $\partial M_\infty(r,t)/\partial t = 0$. For $r > K$ we have $M_\infty(r) = 0$ and for $r = K$ and $r = K - 1$:

$$M_\infty(K) = \frac{\lambda}{2\mu}\frac{L}{K} \tag{A.2}$$

$$M_\infty(K-1) = \frac{\lambda}{\mu}\frac{L}{(K-1)K}\,. \tag{A.3}$$

For $r < K - 1$ we can deduce the following recursion relation:

$$M_\infty(r) = \frac{2}{r}\sum_{s>r} M_\infty(s) \tag{A.4}$$

$$= \frac{2}{r}\left(M_\infty(r+1) + \sum_{s>r+1} M_\infty(s)\right) \tag{A.5}$$

$$= \frac{r+3}{r}\,M_\infty(r+1) \tag{A.6}$$

where we used Eq. (A.4) with $r \to r+1$ to substitute the sum in Eq. (A.5). With this recursion we then compute:

$$M_\infty(r) = \frac{r+3}{r}\frac{r+4}{r+1}\frac{r+5}{r+2}\cdots\frac{K+1}{K-2}\frac{1}{K-1}\frac{L}{K}\frac{\lambda}{\mu}$$

$$= \frac{\lambda}{\mu}\frac{L(K+1)}{r(r+1)(r+2)}\,. \tag{A.7}$$

The stationary $M_\infty(r)$ has clearly a r^{-3} asymptotics for large r. The finite size corrections in the denominator are most relevant for small match length r, a regime of the distribution that is anyway dominated by random matches as described by Eq. (1.39) in the main text.

The differential Eq. (A.1), neglects some contributions, e.g. the loss of matches due to subsequent segmental duplications into a previously duplicated region as well as the additional gain of matches due to the duplication of a sequence that

already contains matches. These two processes occur with probability proportional to the rate λ but compensate each other. Further terms we neglect involve squares and higher powers of rates or lengths much smaller than K and L.

From these considerations we can also deduce that it would be possible to formulate the sequence evolution model such that segmental duplications extend the existing sequence instead of overwriting an already exiting sequence segment. In this case, λK would also be the growth rate of the sequence. However, one has to take special care to define the stationary MLD for the growing system. Only if nucleotides are more frequently changed by mutations than newly generated by segmental duplications, i.e. if $\mu > \lambda K$, a stationary MLD will be reached and the steadily growing version of the model would produce a similar power-law tail.

Appendix B

Non RepeatMasked MLD

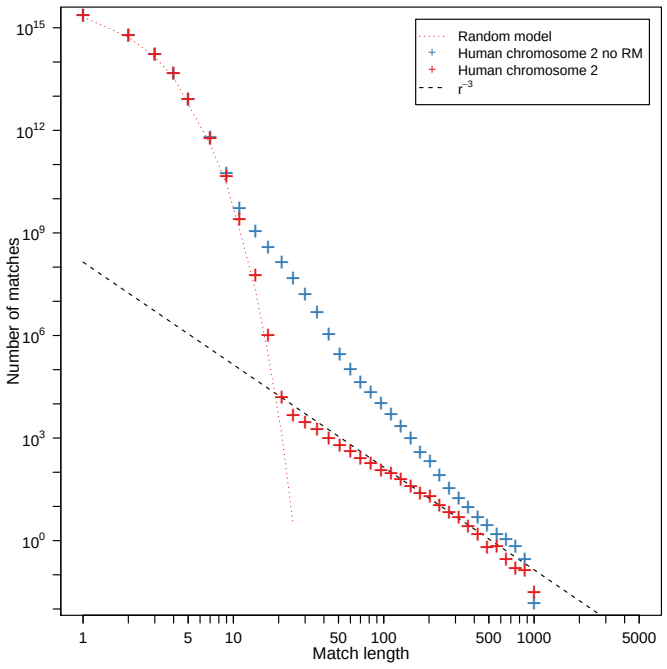

FIGURE B.1: The match length distribution (MLD) computed from the self-alignment of the RepeatMasked human chromosome 2, with (red) and without (blue) RepeatMasking. The red dotted lines represent the distribution obtained when repeating the same experiment on a random iid. sequence with same length and equal nucleotide frequencies. The dashed lines represent the power-law functions L/r^3, where L is the length of the RepeatMasked human Genome (we do not count the Ns). Both MLDs are represented using logarithmic binning, see Section 2.2 for a discussion on this subject. One can see that the power-law behavior is lost when the data are not RepeatMasked. Similar behavior are observed when computing the same experiment on other chromosomes, or on the entire genome.

Appendix C

Article: Statistical Properties of Pairwise Distances between Leaves on a Random Yule Tree

PLOS | ONE

RESEARCH ARTICLE

Statistical Properties of Pairwise Distances between Leaves on a Random Yule Tree

Michael Sheinman[1]*, **Florian Massip**[1,2], **Peter F. Arndt**[1]

1 Max Planck Institute for Molecular Genetics, Berlin, Germany, **2** INRA, UR1077 Unite Mathematique Informatique et Genome, Jouy-en-Josas, France

* mishashe@gmail.com

Abstract

A Yule tree is the result of a branching process with constant birth and death rates. Such a process serves as an instructive null model of many empirical systems, for instance, the evolution of species leading to a phylogenetic tree. However, often in phylogeny the only available information is the pairwise distances between a small fraction of extant species representing the leaves of the tree. In this article we study statistical properties of the pairwise distances in a Yule tree. Using a method based on a recursion, we derive an exact, analytic and compact formula for the expected number of pairs separated by a certain time distance. This number turns out to follow a increasing exponential function. This property of a Yule tree can serve as a simple test for empirical data to be well described by a Yule process. We further use this recursive method to calculate the expected number of the *n*-most closely related pairs of leaves and the number of cherries separated by a certain time distance. To make our results more useful for realistic scenarios, we explicitly take into account that the leaves of a tree may be incompletely sampled and derive a criterion for poorly sampled phylogenies. We show that our result can account for empirical data, using two families of birds species.

Citation: Sheinman M, Massip F, Arndt PF (2015) Statistical Properties of Pairwise Distances between Leaves on a Random Yule Tree. PLoS ONE 10(3): e0120206. doi:10.1371/journal.pone.0120206

Academic Editor: Arndt von Haeseler, Max F. Perutz Laboratories, AUSTRIA

Received: October 10, 2014

Accepted: January 20, 2015

Published: March 31, 2015

Data Availability Statement: All relevant data are within the paper.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

The speciation process in evolution can be regarded as a branching process. One of the simplest stochastic models for a branching process is the so called Yule process [1, 2]. In this model branches are assumed to split with a constant rate and both resulting branches will evolve independently in time. Starting from one branch, a tree will grow, such that the number of leaves on average increases exponentially in time. In a more general version of the Yule tree each branch can also die and get extinct with a constant rate.

Despite its simplicity, many phenomena in different fields of science have been successfully modeled using the Yule process [3, 4]. Particular examples include statistical properties of the number of species in a genus [1], the number of members in protein and gene families [5, 6] and phoneme frequencies in languages [7]. In stochastic modelling of biological evolution, the Yule process is often useful as an instructive null hypothesis [8–11], even when its assumptions are clearly violated.

As an illustrative example of the branching process we present the reconstructed phylogenetic tree of species in the Siilvidae family of birds in the left panel of Fig. 1. The basis of such a reconstructed tree is pairwise distances between individual species. The color-coded matrix of such distances for the species is shown in the right panel of Fig. 1. The statistical properties of such a matrix for a Yule tree is the focus of our article.

Statistical properties of Yule trees have been intensively studied and much is already known. One of the most useful results is the distribution of the number of leaves on a Yule tree [12]. This exact analytical result is widely exploited, in particular, for reconstruction of phylogenetic trees and for estimation of rates of speciation and extinction [10, 11, 13]. Other discrete properties have been studied in Refs. [14–17] as well as properties of the distribution of branch lengths [18, 19].

Often the pairwise distances between all pairs of species in a group of species is the only available information useful for reconstruction of the evolutionary history of the group. For example, in phylogeny reconstruction, one can estimate the pairwise distance in time between two species (twice the time to their last common ancestor) using the molecular clock approach, together with morphological considerations and information about the fossil record [20]. Motivated by observations of mitochondrial DNA sequences with no recombination, the distribution of pairwise distances has been studied in Ref. [21] for a tree with discrete generations and a given number of leaves. In this study, the authors use a sort of mean-field approach, ignoring fluctuations in the number of leaves during the growth of the tree, to derive an approximate formula for the pairwise distances distribution on a tree.

Here we present a general method to derive the distribution of pairwise distances and other statistical properties on a continuous random Yule tree of a certain height with given birth and death rates. Using our method, we obtain exact, analytic, closed, non-recursive and compact formulas for the pairwise distance distribution, the distribution of distances to the closest neighbour, the distance distribution in so-called cherries, as well as a more general formula for the distribution distance to the n-th closest neighbour.

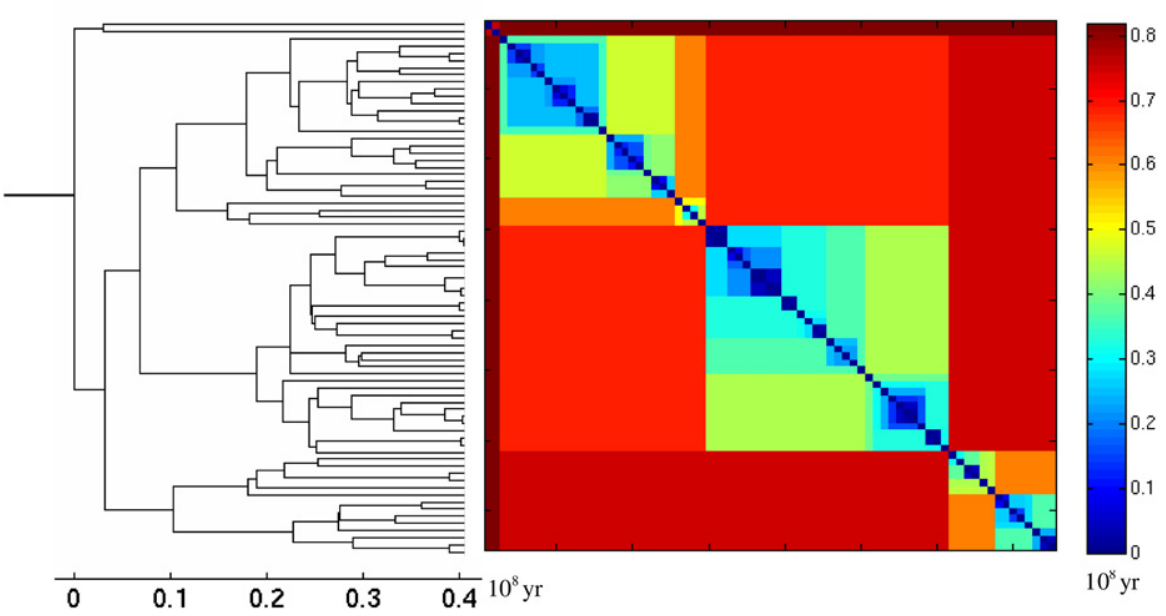

Fig 1. One of the reconstructed trees for the Siilvidae family of species, taken from [28] (left) and its distance matrix (right). The tree includes only the branches which lead to survived and observed leaves.

doi:10.1371/journal.pone.0120206.g001

Often, in biological context, one does not have an access to data about all existing species (i.e. leaves of a phylogenetic tree) [22]. Instead, species are incompletely sampled, or might have been subject to a recent massive extinction event [23]. As long as the extinction of species is random, both scenarios are equivalent on macroevolutionary timescales. In our study, we take the incomplete sampling explicitly into account, which allows us to make statements about the fraction of sampled species, using only the available data.

In the next section we will start with a formal definition of the Yule process and then derive the above mentioned distributions of pairwise distances. For illustrative purposes we also present numerical simulations perfectly matching our expectations. At the end of our article we apply our theoretical consideration to empirical data and analyze the speciation process in two families of birds for which data on speciation times and pairwise distances is available. One advantage of our approach is that we do not need to reconstruct a phylogenetic tree but can solely work with data on pairwise distances.

A Yule tree with constant branching and extinction rates and incomplete sampling of leaves

Definition of the Yule Tree

A Yule tree is defined as follows [1, 2]. At time $t = 0$ there is one individual. As time progresses, this individual can branch and give birth to another individual. In an infinitesimally short time interval $[t, t+dt]$, all individuals can give birth to another one, each with the probability λdt. The probability of an individual to die in the same time interval is μdt. We consider an ensemble of trees of age (height) T, referring to all existing individuals at this time as *leaves*. To make the model more realistic, we assume that due to incomplete sampling (or a short massive extinction event) just before the time T, each leaf is observed with a certain probability $0 \leq \sigma \leq 1$. The described process is illustrated in Fig. 2. We assume that the incompleteness of the sampling is random and ignore possible biases due to different sampling schemes [24].

A Few Useful Results for Random Trees Generated by a Yule Process

Consider a Yule tree with birth rate λ and death rate μ, that have been grown for total time (height) T. In the case where all leaves are sampled ($\sigma = 1$), let $P(M|T, \sigma = 1)$ be the probability that there are M leaves on a tree of age T. Following [25], we can then write the probability that no individual ($M = 0$) survives through to time T as

$$P(M = 0|T, \sigma = 1) = 1 - \frac{\lambda - \mu}{\lambda - \mu e^{-(\lambda-\mu)T}}. \quad (1)$$

For $M > 0$ we have

$$P(M|T, \sigma = 1) = \frac{\lambda - \mu}{\lambda - \mu e^{-(\lambda-\mu)T}} \left[1 - \frac{1 - e^{-(\lambda-\mu)T}}{1 - \frac{\mu}{\lambda} e^{-(\lambda-\mu)T}}\right] \left[\frac{1 - e^{-(\lambda-\mu)T}}{1 - \frac{\mu}{\lambda} e^{-(\lambda-\mu)T}}\right]^{M-1}. \quad (2)$$

We can derive corresponding equations also for the case where species are sampled incompletely. In this case, the probability that no species is observed is

$$P(M = 0|T) = P(0|T, \sigma = 1) + \sum_{m=1}^{\infty} \binom{m}{0} \sigma^0 (1 - \sigma)^{m-0} P(m|T, \sigma = 1) = \frac{e^{\mu T}(\mu - \lambda + \sigma\lambda) - e^{\lambda T}\mu\sigma}{e^{\mu T}(\mu - \lambda + \sigma\lambda) - e^{\lambda T}\lambda\sigma} \quad (3)$$

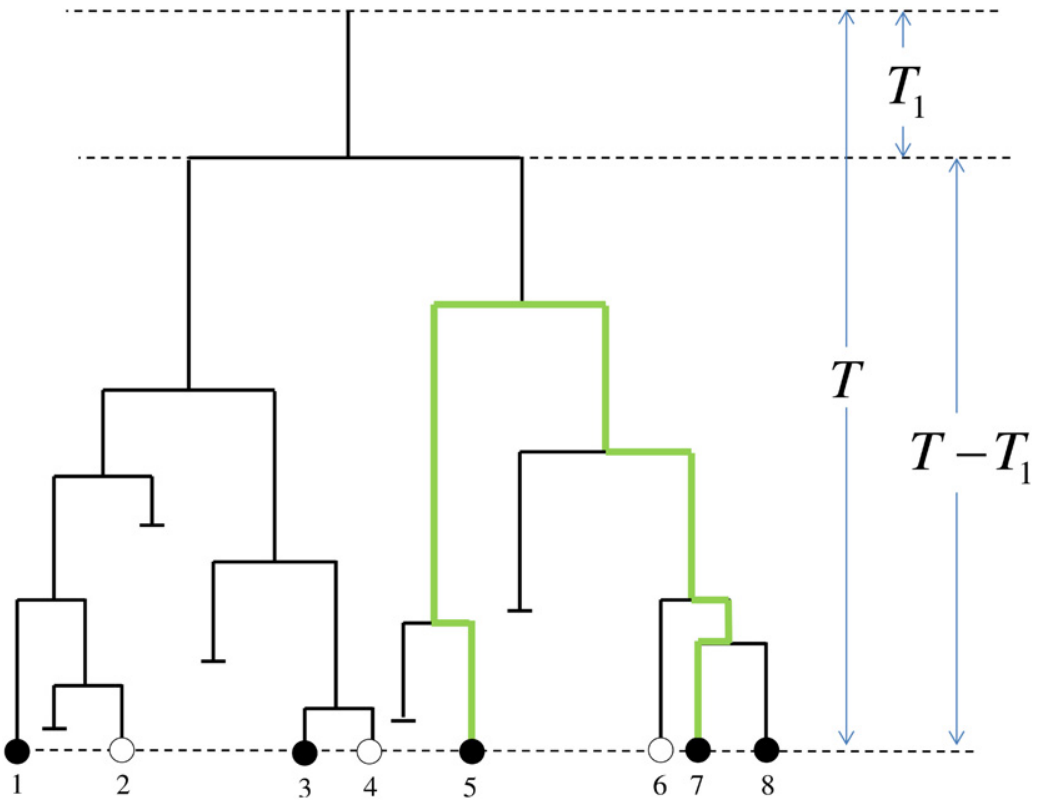

Fig 2. An example of the rooted Yule tree of age *T*. Filled circles (1, 3, 5, 7 and 8) denote observed leaves. Empty circles (2, 4 and 6) denote survived but not observed leaves. Short horizontal lines denotes an extinction event. Long, dashed horizontal lines denote the origin of the tree, the first branching event and the time of sampling the tree, from top to bottom. After the first branching at time T_1 the two resulting subtrees both encompass $M_1 = M_2 = 4$ leaves. However, the number of observed leaves is 2 (leaves 1 and 3) for the left subtree and 3 (leaves 5, 7 and 8) for the right one. The thick green line denotes the pairwise evolutionary distance between the two observed leaves 5 and 7. The horizontal dimension is meaningless. In this example for leaf 1 the first closest observed leaf is 3, the second (as well as the third and the fourth) is 5 (or 7 or 8). The tree has two observed cherry pairs: (1, 3) and (7, 8).

doi:10.1371/journal.pone.0120206.g002

and for $M > 0$

$$P(M|T) = \sum_{m=M}^{\infty} \binom{m}{M} \sigma^M (1-\sigma)^{m-M} P(M|T, \sigma = 1) = \frac{\left[e^{T(\mu-\lambda)} - 1\right]^{M-1} \lambda^{M-1} (\lambda-\mu)^2 \sigma^M e^{MT(\lambda-\mu)}}{\left[\lambda\sigma - \lambda + \mu - \lambda\sigma e^{T(\lambda-\mu)}\right]^{M+1}}. \quad (4)$$

Despite these complicated expressions, the average number of observed leaves in a tree of age T is simply given by

$$\langle M(T) \rangle = \sum_{m=0}^{\infty} m\, P(m|T) = \sigma e^{(\lambda-\mu)T} \quad (5)$$

and the average total number of pairs is

$$\sum_{m=0}^{\infty} \frac{m(m-1)}{2} P(m|T) = \frac{\sigma^2 \lambda}{\lambda - \mu} e^{(\lambda-\mu)T} \left[e^{(\lambda-\mu)T} - 1\right]. \quad (6)$$

The total length of all branches in a Yule tree is given by the integral:

$$\int_0^T \langle M(T)\rangle dt = \int_0^T e^{(\lambda-\mu)t} dt = \frac{1}{\lambda-\mu}\left[e^{(\lambda-\mu)T}-1\right]. \tag{7}$$

To derive a corresponding expression for a a tree reconstructed only from incompletely sampled leaves, we note that the average number of branches at time t with at least one observed descendant at time T is given by

$$\langle M(t,T)\rangle = e^{(\lambda-\mu)t}[1-P(0|T-t,\sigma)]. \tag{8}$$

In the case where $t = T$, we have that $\langle M(T,T)\rangle = \sigma\langle M(T)\rangle$. The average total branch length on the tree of length T excluding the branches which do not lead to an observed leaf is then given by

$$\int_0^T \langle M(t,T)\rangle\, dt = \frac{\sigma e^{T(\lambda-\mu)}}{\mu-\lambda+\sigma\lambda} \ln\frac{\lambda\sigma+(\lambda-\sigma\lambda-\mu)e^{T(\mu-\lambda)}}{\lambda-\mu}. \tag{9}$$

In the limit of no extinction, $\mu \to 0$, and exhaustive sampling, $\sigma \to 1$, Equation (9) is identical to Equation (7). We turn now to calculations of the statistical properties of pairwise distances, using the above formulas.

The Distribution of Pairwise Distances

In a biological context the available data often consist of the pairwise distances separating any pair in a group of species. Commonly these distances are used to reconstruct a phylogenetic tree representing the evolutionary history of a group of species. From such a tree one can then try to estimate rates of speciation and extinction [10, 11]. Here we propose another approach of analysing such data on pairwise distances circumventing the reconstruction of a phylogenetic tree, provided that the pairwise distances between the leaves are properly estimated.

Let $N(t|T)dt$ be the average number of pairs of leaves on a tree of length (evolution time) T, separated by a time distance in the interval $[t, t+dt]$, i.e. their last common ancestor lived in the time interval $[T-t/2-dt/2, T-t/2]$. Now consider the branching process as illustrated in Fig. 2. The first branching happened at time T_1 and the two resulting subtrees encompass, say, M_1 and M_2 leaves, respectively. In this situation one can derive the following recursion relation

$$N(t|T) = [2N(t|T-T_1) + \sigma^2 M_1 M_2\, \delta(t-2(T-T_1))I(0 \le t \le 2T)]e^{-\mu T_1} \tag{10}$$

where the first part in the summation on the right hand side counts the pairs inside each of the two subtrees and the second one counts the pairs between them. The common multiplicative factor, $e^{-\mu T_1}$, expresses the probability that the first branch survives to the time T_1 (otherwise, $N(t|T) = 0$). The function I is the indicator function, defined by:

$$I(\text{condition}) = \begin{cases} 1 & \text{if condition holds} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

and $\delta(x)$ is the Dirac delta function. Averaging over M_1, M_2 (using Equations (3, 4) with time $T-T_1$) and then T_1, which follows an exponential distribution with mean $1/\lambda$, one obtains:

$$N(t|T) = 2\lambda \int_0^\infty N(t|T-T_1)e^{-(\lambda-\mu)T_1} dT_1 + \frac{\sigma^2\lambda}{2} e^{\lambda t} e^{-(\lambda+\mu)(T-t/2)} I(0 \le t \le 2T). \tag{12}$$

In Laplace space one gets:

$$N(t|S) = 2\lambda \frac{N(t|S)}{S+\lambda+\mu} + \frac{\sigma^2\lambda}{2}\frac{e^{\lambda t - St/2}}{S+\lambda+\mu}, \tag{13}$$

where S is the Laplace conjugate variable of T. Solving and inverting the Laplace transform one finally gets the solution:

$$N(t|T) = \frac{\sigma^2\lambda}{2}\, e^{(\lambda-\mu)T}\, e^{(\lambda-\mu)t/2} \tag{14}$$

for $0 \leq t \leq 2T$ and zero otherwise. Fascinatingly, this distribution is a simple exponential function in t. The distribution is cut off at $t = 2T$ because in a tree of age T two leaves cannot be separated by a time larger than $2T$. In Fig. 3(a) we show this distribution of pairwise distances for several parameter values together with results of numerical simulations, which match perfectly our theoretical expectations. This result, applied for trees of DNA sequences can account for statistics of exact sequence matches in genomes of eukaryotes [26].

One can also derive the same result (14) using the following simple arguments. Pairs, separated by a time in the interval $[t, t+dt]$, branched at the time interval $[T-t/2-dt/2, T-t/2]$. The average number of branches in this interval is given by $\lambda e^{(\lambda-\mu)(T-t/2)}\, dt/2$. The average number

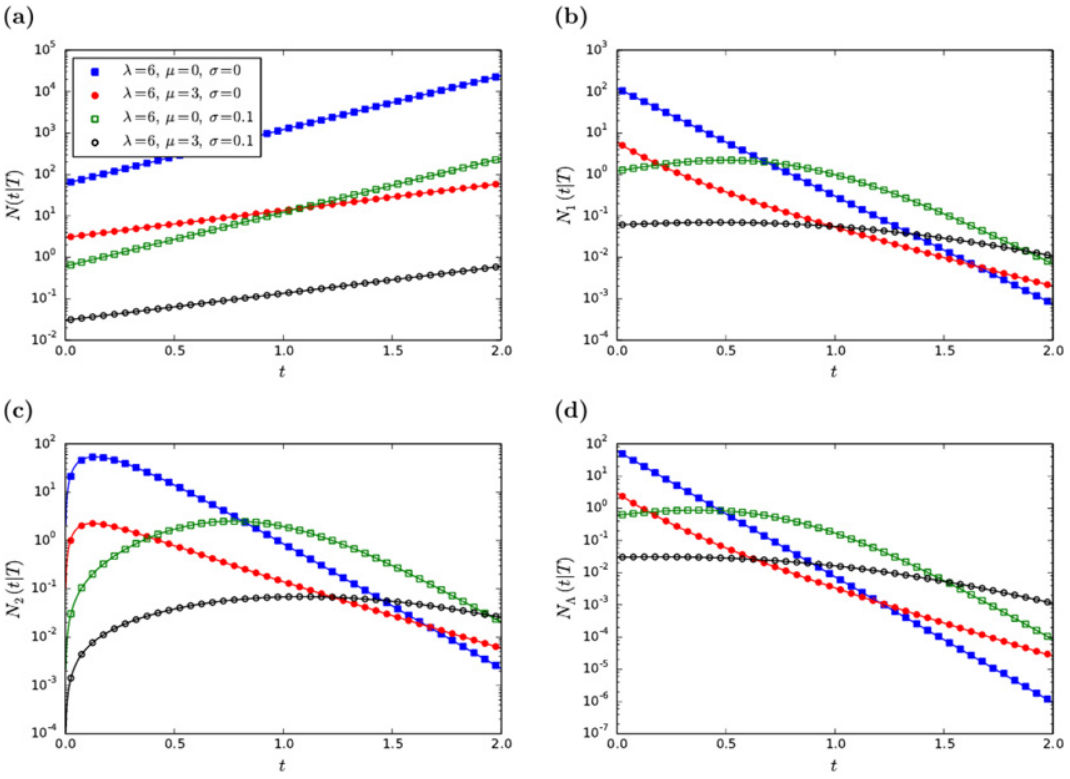

Fig 3. Comparison of the analytic results with numerical simulations. Markers indicate numerically obtained data using the following parameters set. $T = 1$, $\lambda = 6$, $\mu = 0$ or 3 (circles or squares) and $\sigma = 1$ or 0.1 (empty or filled symbols). Lines represent the analytic formulas. (a) Density of number of pairs separated by a certain time, t. Lines were obtained using Equation (14). (b) Density of number of leaves separated by a certain time, t with their closest leaf. Lines were obtained using Equation (17) or Equation (20) with $n = 1$. (c) Density of number of leaves separated by a certain time, t with their next-closest leaf. Lines were obtained using Equation (33) or Equation (20) with $n = 2$. (d) Density of number of cherries separated by a certain time, t. Lines were obtained using Equation (21).

doi:10.1371/journal.pone.0120206.g003

of observed pairs from a branch at this time is given by $(\sigma e^{(\lambda-\mu)t/2})^2$. Multiplying the two factors one gets Equation (14). However, for other quantities, derived below, the recursive equation approach is more effective.

The Distribution of the Minimal-Distance to Other Leaves

Using the recursive method from the previous Section one can also compute other interesting quantities. For instance, in certain situations, the distance separating a leaf to its most closely relative may be estimated more precisely than its distance to other leaves in the tree. Thus, we might be interested in $N_1(t|T)dt$—the average number of leaves on the tree of age T, separated by the time distance between t and $t+dt$ from their most closely related leaf. Interestingly, calculating this quantity lets us make certain statements on the value of the sampling rate σ.

To calculate this distribution, we can again write a recursion relation, assuming that the first branching occurred at time T_1. In this case one gets the distribution of the minimal distance time in the form

$$N_1(t|T) = \{2N_1(t|T-T_1) + 2P(1|T-T_1)[1-P(0|T-T_1)]\delta(t-2(T-T_1))I(0 \le t \le 2T)\}e^{-\mu T_1}, \quad (15)$$

where $P(M|T)$ is the probability to observe M leaves after time T, as computed in Equations (3) and (4). In contrast to the recursion relation for the distribution of all pairwise distances, we count a branching point only if $M_1 = 1$ and $M_2 > 0$ or $M_1 > 0$ and $M_2 = 1$, as expressed by the product $2P(1|T–T_1)[1–P(0|T–T_1)]$ in Equation (15).

Averaging Equation (15) over T_1, one gets:

$$N_1(t|T) = 2\lambda \int_0^\infty N_1(t|T-T_1)e^{-(\lambda+\mu)T_1}dT_1 + \frac{e^{-(\lambda+\mu)T+(3\lambda/2+\mu)t}\lambda(\lambda-\mu)^3\sigma^2}{\left[e^{\frac{\lambda t}{2}}\lambda\sigma - e^{\frac{\mu t}{2}}(\mu-\lambda+\sigma\lambda)\right]^3} I(0 \le t \le 2T). \quad (16)$$

The solution of this equation is given by

$$N_1(t|T) = \frac{e^{\frac{\lambda t}{2}+\lambda T+\mu t-\mu T}\lambda(\lambda-\mu)^3\sigma^2}{\left[e^{\frac{\lambda t}{2}}\lambda\sigma - e^{\frac{\mu t}{2}}(\mu-\lambda+\sigma\lambda)\right]^3} \quad (17)$$

for $0 \le t \le 2T$ and 0 otherwise. Results of numerical simulations perfectly match our theoretical expectations (see Fig. 3(b)). Interestingly, the function $N_1(t|T)$ from Equation (17) possesses a maximum only if

$$\sigma < \frac{1}{3}\left(1-\frac{\mu}{\lambda}\right) \le \frac{1}{3} \quad (18)$$

and the position of the maximum

$$t^1_{\max} \equiv \frac{2}{\lambda-\mu}\ln\frac{\lambda(1-\sigma)-\mu}{2\lambda\sigma} \quad (19)$$

is in the range $[0, 2T]$. This result is useful for a quick estimation of the data completeness. In particular, a maximum in the distribution of the minimal distance implies that the sampling of the considered tree is not complete and $\sigma < 1/3$.

By similar arguments we can also derive expressions for the distributions of second minimal distances, $N_2(t|T)$ (see Appendix) and of the n-th minimal distance $N_n(t|T)$ (see Appendix) to

other leaves. The latter quantity is computed to be

$$N_n(t|T) = \frac{n(1+n)(\mu-\lambda)^3\sigma(\lambda\sigma)^n}{2} \frac{\left[e^{\frac{1}{2}t(\mu-\lambda)}-1\right]^{n-1} e^{\frac{t\mu}{2}+T\lambda+t\mu-T\mu}}{\left[e^{\frac{t\mu}{2}}(\mu-\lambda+\sigma\lambda)-e^{\frac{t\lambda}{2}}\lambda\sigma\right]^{n+2}} \tag{20}$$

for $0 \leq t \leq 2T$ and 0 otherwise. In Appendix we also calculate the distribution of distances in "cherries". Cherries are adjacent pairs of leaves, such that they are reciprocal closest neighbors to each other (see Fig. 2 for illustration of cherries):

$$N_\Lambda(t|T) = \frac{\lambda(\lambda-\mu)^4\sigma^2}{2} \frac{e^{\frac{t\lambda}{2}+T\lambda+\frac{3t\mu}{2}-T\mu}}{\left[e^{\frac{t\mu}{2}}(\mu-\lambda+\sigma\lambda)-e^{\frac{t\lambda}{2}}\lambda\sigma\right]^4} \tag{21}$$

for $0 \leq t \leq 2T$ and 0 otherwise. The function $N_\Lambda(t|T)$ from Equation (21) possesses a maximum only if

$$\sigma < \frac{1}{4}\left(1-\frac{\mu}{\lambda}\right) \leq \frac{1}{4} \tag{22}$$

and the position of the maximum

$$t^\Lambda_{\max} \equiv \frac{2}{\lambda-\mu} \ln \frac{(1-\sigma)\lambda-\mu}{3\lambda\sigma} \tag{23}$$

is in the range $[0, 2T]$. This result is useful for a quick estimation of the data completeness. In particular, a maximum in the distribution of the distance between cherries implies that the sampling of the considered tree is not complete and $\sigma < 1/4$.

For illustration purposes we show the distributions for the second minimal distance in Fig. 3(c) and, for cherries, in Fig. 3(d).

Beyond the Averages

Above results are average expectations. For instance, in The Distribution of Pairwise Distances Section we derive $N(t|T)$, defined as the *average* density number of pairs, separated by a certain time distance t, on a tree of length T. The average is over many realizations, say S many, of the Yule trees with a given set of parameters λ, μ, σ and T. Namely,

$$N(t|T) = \langle N^s(t|T)\rangle_s = \lim_{S\to\infty} \frac{1}{S}\sum_{s=1}^{S} N^s(t|T), \tag{24}$$

where $N^s(t|T)$ is the density number of pairs separated by a time distance in the interval $[t, t+dt]$ in an individual sample tree number s. In reality one often possesses information only about one specific tree $s = 1$, i.e. $N^1(t|T)$. Therefore, we are interested not only in the derived averages of $N(t|T)$, $N_n(t|T)$, $N_\Lambda(t|T)$ etc. but also their distributions in finite time intervals. The last becomes especially important in the maximum likelihood fitting and model testing. In the discussion below we refer to the distribution of the number of pairs separated by a certain time, $N^1(t|T)$. However, the same arguments can be applied to other quantities, like the n-th minimal distance or the distance in cherries, which we mention above.

Consider an infinitesimal (in practice very small) interval, $[t, t+dt]$, such that $N(t|T)dt \ll 1$. The number of pairs $N^1(t|T)dt$ in this interval is distributed with the mean $N(t|T)dt$. However, in the considered small bin limit, the mean does not represent well the typical value because the distribution of $N^1(t|T)dt$ is not well peaked but possesses a very small probability of having any positive value, while probability of having zero is almost one (see Appendix).

Pairs separated by the time in the interval $[t, t+dt]$ branched at the time interval $[T-t/2-dt/2, T-t/2]$. The probability to have a branch in this interval is given by $\lambda e^{(\lambda-\mu)(T-t/2)}\, dt/2$. Given that there is a branching point in this interval it can lead to different number of leaves. The probability that no observed pairs survive from this branching is given by $1-[1-P(0|t/2)]^2$, where $P(M|T)$ is the probability to observe M leaves on a tree of age T and is given in Equations (3, 4). Therefore, the probability that there are no observed pairs separated by the time in the interval $[t, t+dt]$ is given by

$$\Pr\left(N^1(t|T)dt = 0\right) = 1 - \lambda e^{(\lambda-\mu)(T-t/2)} dt/2\{1-[1-P(0|t/2)]^2\}. \quad (25)$$

In sum, in the small bin limit it is convenient to break the full distribution in two distributions: One comprising only the peak at zero and a second representing all samples with $N^1(t|T)\, dt \neq 0$. The total average can be broken as follow:

$$N(t|T)dt = 0 \times \Pr\left(N^1(t|T)dt = 0\right) + \tilde{N}(t|T)dt \times [1 - \Pr\left(N^1(t|T)dt = 0\right)]. \quad (26)$$

Here $\tilde{N}(t \mid T)$ is the average of $N^1(t|T)$ over the tree realizations with $N^1(t|T) > 0$. It can be computed to be:

$$\tilde{N}(t|T) = \lim_{S\to\infty} \frac{\sum_{s=1}^{S} N^s(t|T)}{\tilde{S}(t)} = \frac{N(t|T)}{1 - \Pr\left(N^1(t|T)dt = 0\right)} = \frac{1}{dt}\left(1 + \sigma\lambda \frac{e^{\frac{\lambda-\mu}{2}t} - 1}{\lambda - \mu}\right)^2, \quad (27)$$

where $\tilde{S}(t) = \sum_{s=1}^{S}\left[1 - \delta_{N^s(t|T),0}\right]$ is the number of samples with $N^1(t|T) > 0$. Since, $1-\Pr\left(N^1(t|T)dt = 0\right) \ll 1$, the value of $N(t|T)dt$ is not representative of the expected empirical average of $N^1(t|T)dt$ for finite S and, in particular, $S = 1$. However, the value of $\tilde{N}(t \mid T)$, derived above (see Equation (27)), is representative of the expected empirical average of positive values of $N^s(t|T)dt$. We illustrate this in Fig. 4

Constrains on the sampling fraction

One can easily see that all the derived above results do not depend explicitly on the parameters λ, μ and σ, but only on their combinations: $\lambda-\mu$ and $\sigma\lambda$. Therefore, one cannot estimate the sampling fraction, σ, based on fitting the empirical data to the derived formulas (see examples in the next Section). The same loss of information in reconstructed trees was reported, based on an analysis of the density of bifurcation times in the reconstructed tree [27].

However, the information about the values of λ, μ and, most intriguingly, σ is not lost completely. For instance, observing a maximum in the distribution of the minimal distances one can deduce that $\sigma < 1/3$ (see Equation (18)). Observing a maximum in the distribution of the distances between cherries one can deduce that $\sigma < 1/4$ (see Equation (23)). It is of an interest to construct other distributions which, possessing a maximum, provide information about the value of the sampling fraction, σ.

Consider an average density of pairs of leaves with the following property. Given that the first (second) leaf of the pair has a nearest neighbor at a distance (if a leaf is alone in the tree we define the distance to its nearest neighbor as twice the height of the tree) t_1 (t_2) the quantity $\min(t_1, t_2)$ is given by t. We denote this density by $N_{\min 2}(t|T)$. The recursive equation for this quantity is given for a given time of first bifurcation, T_1 by

$$N_{\min 2}(t|T) = \left\{2N_{\min 2}(t|T-T_1) + 2\left[\sigma e^{(\lambda-\mu)(T-T_1)} - \int_0^t N_1(t'|T-T_1)dt'\right] N_1(t|T-T_1)\right\} e^{-\mu T_1} \quad (28)$$

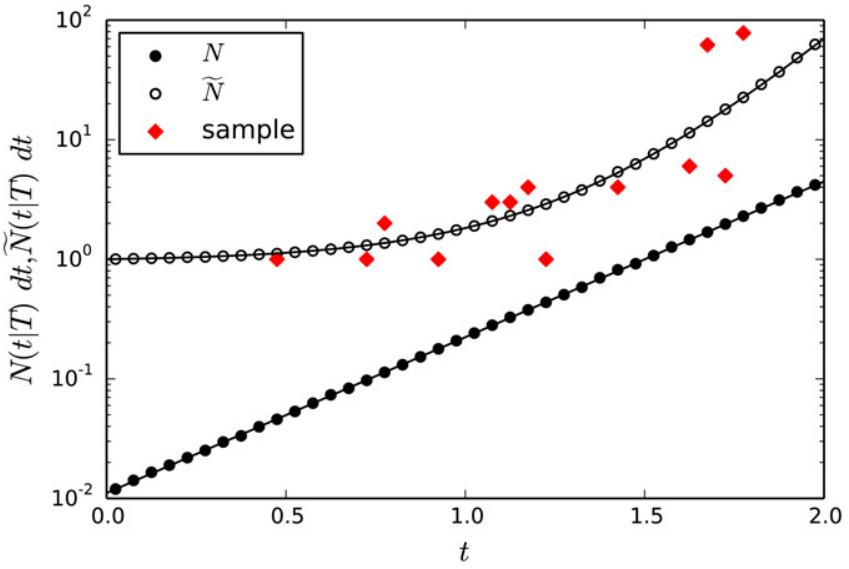

Fig 4. The benefit to use $\tilde{N}(t|T)$ instead of $N(t|T)$ to estimate the parameters of the evolution process in a case of a small dataset. In this plot $T = 1$, $\lambda = 11$, $\mu = 5$, $\sigma = 0.01$ and $dt = 0.005$. After average over many samples ($S \sim 10^6$ in this particular case) empirical averages of both $N(t|T)$ (full circles) and $\tilde{N}(t \mid T)$ (open circles) converge nicely to the analytic formulas. The last are given in Equations (14) and (27), respectively, and are denoted by the lines in the figure (see the legend). However, for a single random tree, $S = 1$, the values of $N^1(t|T)$ (diamonds) are highly dispersed (most intervals show zero counts and do not show up in the semilogarithmic plot), such that their fit to the analytic formula of $N(t|T)$ is not expected to lead to a good estimation of the model's parameters. In contrast, the values of $N^1(t|T)$, ignoring the bins where $N^1(t|T) = 0$, are well distributed around $\tilde{N}(t \mid T)$, although in this example the tree possesses only 19 observed leaves, such that the data is very poor (only 171 pairs in total).

doi:10.1371/journal.pone.0120206.g004

After average over T_1 the solution is given by

$$N_{\min 2}(t|T) = \frac{2\lambda^2\sigma^3(\lambda-\mu)^4 e^{T(\lambda-\mu)}}{3\lambda-\mu} \frac{2\lambda e^{t(\lambda-\mu)} + (\lambda-\mu)e^{-\frac{1}{2}t(\lambda+\mu)} + (\mu-3\lambda)e^{\frac{1}{2}t(\lambda-\mu)}e^{T(\lambda-\mu)}}{\left[\lambda\sigma - \lambda + \mu - \lambda\sigma e^{\frac{1}{2}t(\lambda-\mu)}\right]^5}. \quad (29)$$

This function possesses a maximum only if

$$\sigma < \frac{1}{5}\left(1 - \frac{\mu}{\lambda}\right) \leq \frac{1}{5} \quad (30)$$

Therefore, observing a maximum in the distribution of the minimal distance to the closest neighbors between two leaves one can deduce that $\sigma < 1/5$. Using our recursive method one can calculate different distributions (say, the minimal distance to the closest neighbor among three leaves etc.) which, exhibiting a maximum, provide direct information about an upper limit on the sampling fraction.

Comparison of the derived results to empirical data

In this Section we demonstrate the relevance of the obtained analytic formulas to empirical data, studying the pairwise distances between species in families of the evolutionary tree. For comparison with the derived results we choose $N(t|T)$, $N_n(t|T)$ with $n = 1, 2, 3, 4$ and $N_\Lambda(t|T)$. The results are presented in Fig. 5 for the Siilvidae family of birds (see one of the reconstructed trees for this family and its distance matrix in Fig. 1) and for the Tyrannidae family of birds in

Fig. 6. For every family we analyze Bayesian sampling of 1000 trees downloaded from the database [28]. Namely, we collect pairwise distances, n-minimal distances and distances between cherries of all 1000 trees and plot the histograms of these distances (with the y-axis divided by 1000) in Figs. 5 and 6. We fit all the points in a figure using the iterative reweighted least squares algorithm [29] in Matlab. Unfortunately, the explicit dependencies on λ and μ in Equations (14, 20, 21) are insufficient to estimate all parameters. Instead one can estimate from the fit only the effective growth rate, $\lambda-\mu$ and $\lambda\sigma$. The value of σ can be obtained assuming a certain ratio μ/λ. In the captions of Figs. 5 and 6 we present the obtained estimates for σ for different assumptions about the ratio μ/λ.

Over all, the fits to empirical data look satisfactory and result in a reasonable set of parameters, which roughly agree with the ones given in [28]. This indicates that certain statistical properties of speciation can be well captured by a simple Yule process. However, in some cases, deviations can be observed. For example, for the Sylviidae family the pairwise distances distribution deviates from the prediction for $t > 30$ Myr, while for the Tyrannidae family we observe a clear deviation for distances around 55 Myr in all our estimates. This indicates a massive radiation event in the considered family of birds around 27.5 Myr ago, as already reported in [28], or other violation of the Yule process assumptions.

Interestingly, we can state that the Sylviidae family of birds is currently not well sampled. In fact, the estimator for the upper limit of the sampling fraction σ is 30% (see Fig. 5).

Summary and concluding remarks

In this paper we present a novel method to calculate statistical properties of Yule trees. The method is based on a recursive equations which can be solved using the Laplace transform. We demonstrate the strength of our method deriving formulas for (*i*) average number of pairs separated by a certain time (Equation (14)), (*ii*) the number of most closely related pairs separated by a certain time (Equation (17)), (*iii*) the number of next-most closely related pairs separated by a certain time (Equation (33)), (*iv*) the number of n-most closely related pairs separated by a certain time (Equation (20)) and (*v*) the number of cherries separated by a certain time (Equation (21)).

Our results can be compared to empirical data using only the information about pairwise distances between leaves of a considered tree. We assume that the estimation of the pairwise distances is precise enough. If the distances are estimated using genetic divergence, this assume that the molecular clock reflect adequately the real time distance. If this holds the reconstruction of the tree structure is not required. This is a particular strength of our method because the reconstruction of such trees for a large number of leaves is sometimes problematic. In such cases one often considered a posterior distribution of trees which is generated by Bayesian sampling [30, 31]. Such a distribution of trees can still be easily analyzed using our method, based on recursive equations. Analyzing such ensembles of trees we use only their distance matrices.

We demonstrate the relevance of our results to statistical properties of pairwise evolutionary time distances between biological species. We find that in some cases the speciation process is well described by the Yule model. Significant deviations from the derived distributions are expected to be indicative for massive extinction or radiation events. In the case where the assumptions of the Yule process are justified, we expect our results to be useful for estimation of the incompleteness of the data sampling, i.e. the fraction of observed leaves out of all existing leaves, σ. However, similarly to the method developed in Ref. [11], all the derived results depend only on three parameters: $\lambda-\mu$, $\lambda\sigma$ and $\sigma e^{(\lambda-\mu)T}$. Therefore, even knowing those *three* parameters one cannot estimate the values of the *four* unknown parameters: the rates λ, μ, the height of the tree, T and the sampling fraction, σ, without an additional assumption about one

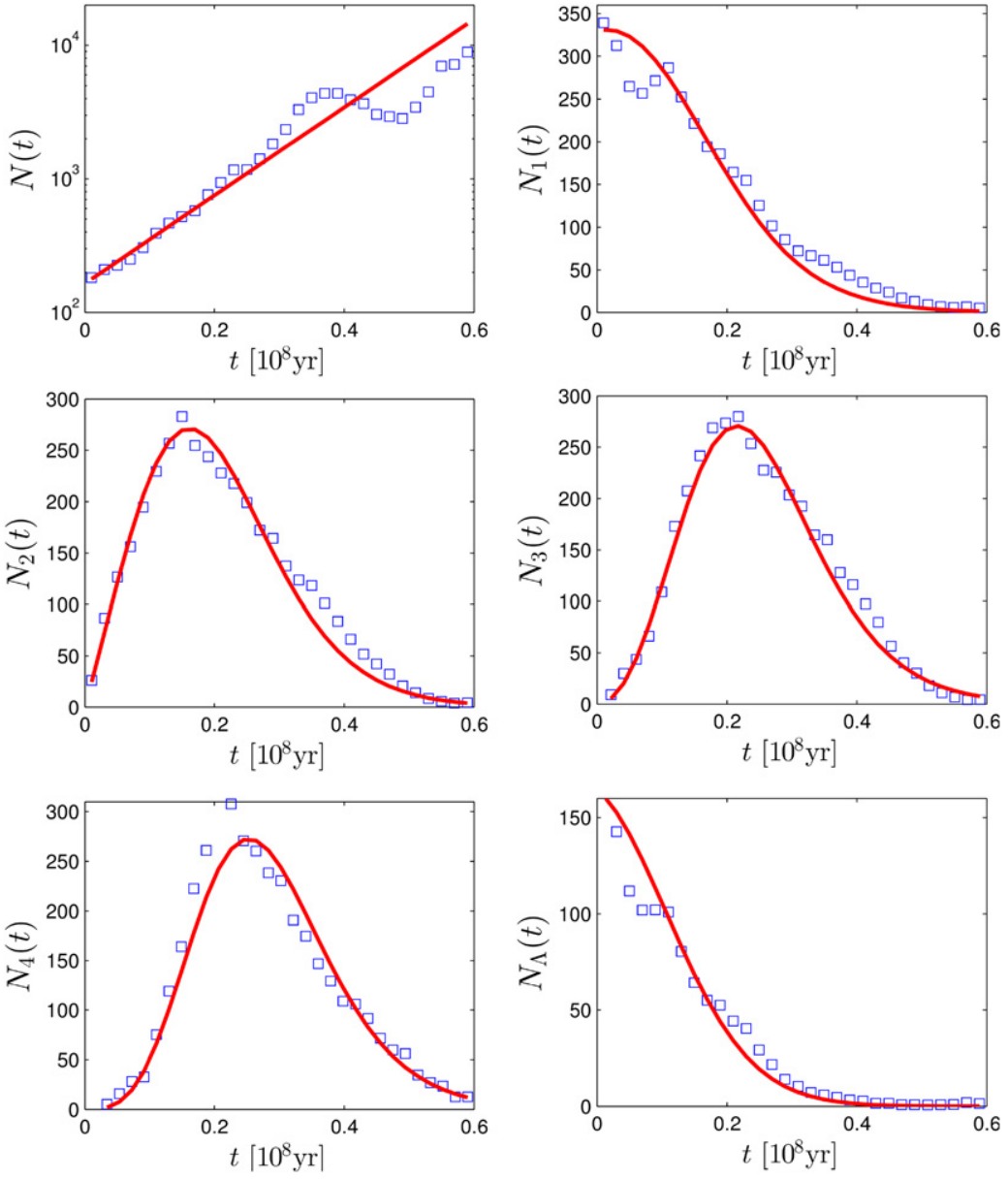

Fig 5. Comparison of analytic predictions to the pairwise distances data of Sylviidae family with $M = 75$ species taken from the database [28] with $t \leq 0.6 \times 10^8$Myr. The markers represent the empirical data, while the lines represent the analytic formulas with fitted parameters. (a) Pairwise distance distribution. (b) Minimal distance distribution.(c-e) *n*-minimal distance distribution. (d) Cherries distance distribution. The lines are based on following set of parameters: $\lambda-\mu = 15.2 \times 10^{-8}\mathrm{yr}^{-1}$ and $\lambda\sigma = 4.6 \times 10^{-8}\mathrm{yr}^{-1}$. For $\mu = 0, 0.2, 0.4, 0.6, 0.8 \times \lambda$ this corresponds respectively to $\sigma = 0.3, 0.24, 0.18, 0.12, 0.06$.

doi:10.1371/journal.pone.0120206.g005

of these parameters, for instance the fraction μ/λ. After estimation of $(\lambda-\mu)$ and $(\lambda\sigma)$ one can get an upper bound for the sampling fraction in the form (note that $\mu \geq 0$)

$$\sigma \leq \frac{(\sigma\lambda)}{(\lambda-\mu)}. \tag{31}$$

If the death rate is known to be much smaller than the birth rate, $0 \leq \mu \ll \lambda$, the upper bound is expected to be a good estimate for σ.

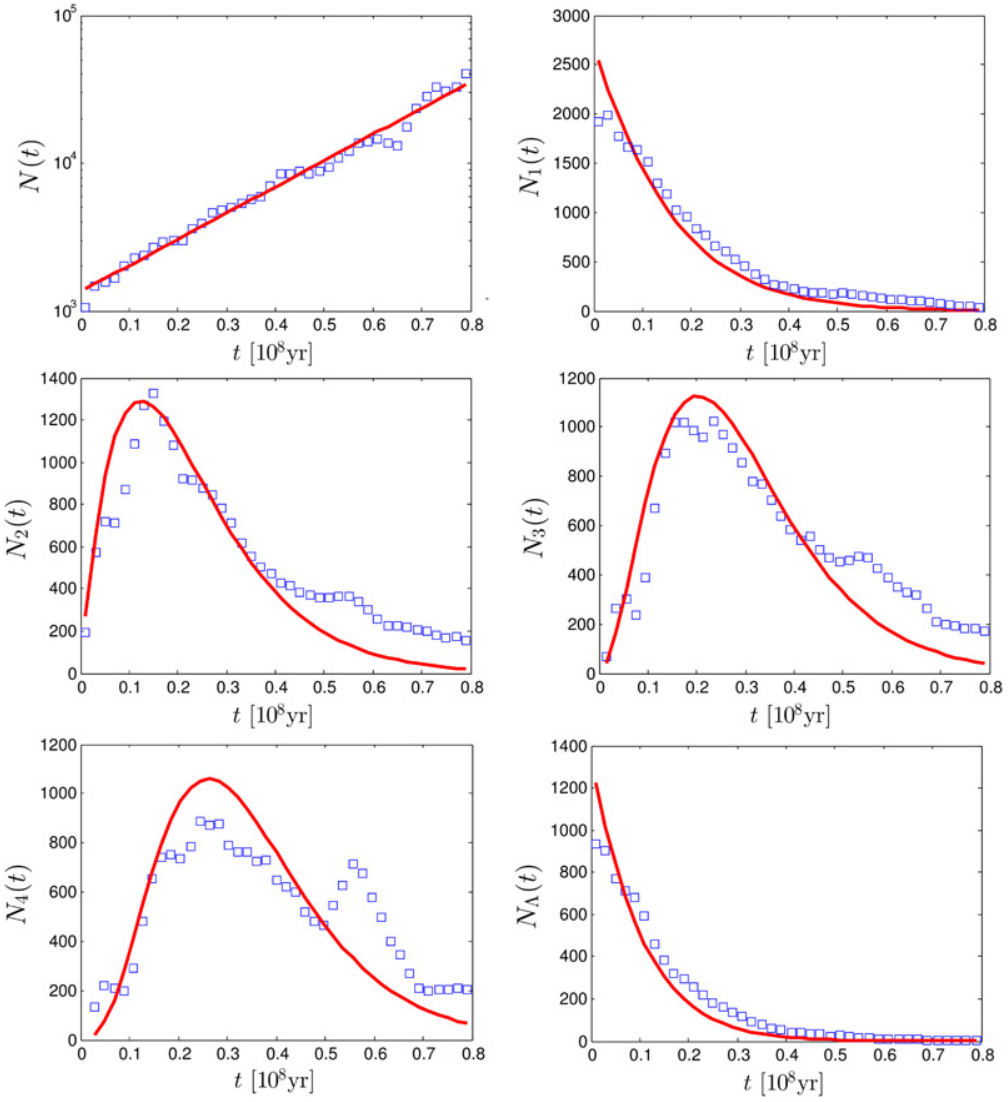

Fig 6. Comparison of analytic predictions to the pairwise distances data of Tyrannidae family with $M = 460$ species taken from the database [28] with $t \leq 0.8 \times 10^8$Myr. The markers represent the empirical data, while the lines represent the analytic formulas with fitted parameters. (a) Pairwise distance distribution. (b) Minimal distance distribution. (c-e) n-minimal distance distribution. (d) Cherries distance distribution. The fit is performed for all points in the figure with $t \leq 0.5$ to avoid clear break down of the Yule tree assumptions for larger distances (see text). The lines are based on following set of parameters: $\lambda - \mu = 8 \times 10^{-8}\text{yr}^{-1}$ and $\lambda\sigma = 6.4 \times 10^{-8}\text{yr}^{-1}$. For $\mu = 0, 0.2, 0.4, 0.6, 0.8 \times \lambda$ this corresponds respectively to $\sigma = 0.8, 0.64, 0.48, 0.32, 0.16$.

doi:10.1371/journal.pone.0120206.g006

If it is known that the sampling is perfect, $\sigma = 1$, one can estimate both the birth and the death rate. However, in contrast to Ref. [11], the method presented here does not require the reconstruction of the tree, but is solely based on statistical properties of pairwise distances between the leaves of the tree.

In the general case, one can get an upper limit for the sampling fraction and a lower limit for the birth rate by setting $\mu/\lambda = 0$. These bounds are expected to be useful for analysis of exponentially growing trees. Such trees can appear in phylogeny when analyzing the evolution of taxa, but also in population genetics, for instance, when considering an exponentially growing sub-population under the influence of a positive selection.

Appendix

Simulation details

To simulate Yule process for the generation of phylogenetic trees we use a Kinetic Monte Carlo algorithm. For a given birth rate λ, death rate μ, and sampling fraction σ, the system is initiated with one "alive" lineage $M = 1$ at time $t = 0$. The system is then iteratively propagated to the time $t = T$. In each iterative step one alive lineage is chosen at random and either either split into two alive lineages (with probability $\lambda/(\lambda+\mu)$) or killed (with probability $\mu/(\lambda+\mu)$). In each step the time is incremented by an amount Δt that is exponentially distributed with mean $1/(M(\lambda+\mu))$, where M is the number of alive lineages. After the time $t = T$ has been reached, alive lineage are kept in the set of sampled leaves with probability σ.

During the whole simulation the complete tree—especially information about all branching points and branching times—are kept in memory. This way the distribution of pairwise distances or other quantities described in the text can easily be computed. To obtain the mean of such distributions we usually generated at least 10^6 trees and computed the averages.

Second-minimal-distance distribution

Let $N_2(t|T)dt$ be the average number of leaves on the tree of length T, separated by the time distance t from their second-most closely related leaf. Then, if the first branching occurs at time T_1 and the two resulting subtrees possess M_1 and M_2 leaves, respectively, one gets the distribution of the minimal distance time in a form

$$\begin{aligned} N_2(t|T) = \; & 2N_2(t|T-T_1)e^{-\mu T_1} \\ & + 2[2P(2|t/2)(1-P(0|t/2)) + P(1|t/2)(1-P(0|t/2)-P(1|t/2))] \\ & \times \delta(t-2(T-T_1))I(0 \le t \le 2T)e^{-\mu T_1}. \end{aligned} \tag{32}$$

After average over T_1 and solving the resulting equation one obtains

$$N_2(t|T) = \frac{3\lambda^2(\lambda-\mu)^3\sigma^3\left(e^{\frac{t\lambda}{2}} - e^{\frac{t\mu}{2}}\right)}{\left[e^{\frac{t\mu}{2}}(\mu-\lambda+\sigma\lambda) - e^{\frac{t\lambda}{2}}\lambda\sigma\right]^4} e^{\frac{t\lambda}{2}+T\lambda+t\mu-T\mu} \tag{33}$$

for $0 \le t \le 2T$. Similarly, one can obtain any third-minimal distance distribution forth- etc. The general formula for the n-minimal-distance distribution is calculated in the following.

n-minimal-distance distribution

Let $N_n(t|T)dt$ be the average number of leaves on the tree of length T, separated by the time distance t from their n-most closely related leaf. This notation means that 1-most closely related leaf is the closest one, 2-most closely related leaf is the second-most closest one etc. Then, if the first branching happens at time T_1 and the two resulting subtrees possess M_1 and M_2 leaves, respectively, one gets the distribution of the minimal distance time in a form

$$\begin{aligned} N_n(t|T) = \; & 2N_n(t|T-T_1)e^{-\mu T_1} \\ & + 2[nP(n|t/2)P_>(0|t/2) + (n-1)P(n-1|t/2)P_>(1|t/2) + \ldots + P(1|t/2)P_>(n-1|t/2)] \\ & \times \delta(t-2(T-T_1))I(0 \le t \le 2T)e^{-\mu T_1} \\ = \; & \left[2N_n(t|T-T_1) + 2\delta(t-2(T-T_1))I(0 \le t \le 2T)\sum_{k=1}^{n} k\,P(k|t/2)P_>(n-k|t/2)\right]e^{-\mu T_1}. \end{aligned} \tag{34}$$

Here

$$P_{>}(k|T) = \frac{\sigma^{k+1}(\mu-\lambda)\lambda^k\left[e^{T(\mu-\lambda)}-1\right]^k\left(e^{T\lambda}\lambda - e^{T\mu}\mu\right)^k e^{T\lambda}}{\left[e^{T\mu}(\mu-\lambda+\sigma\lambda) - e^{T\lambda}\lambda\sigma\right]^{k+1}\left[\lambda - e^{T(\mu-\lambda)}\mu\right]^k} \tag{35}$$

is the probability to observe more than k leaves on a tree of age T and $P(n|T)$ is given in Equations (3, 4). After average over T_1 and solving the resulting equation one obtains

$$N_n(t|T) = \frac{n(1+n)(\mu-\lambda)^3\sigma(\lambda\sigma)^n}{2}\frac{\left[e^{\frac{1}{2}t(\mu-\lambda)}-1\right]^{n-1}e^{\frac{nt\lambda}{2}+T\lambda+t\mu-T\mu}}{\left[e^{\frac{t\mu}{2}}(\mu-\lambda+\sigma\lambda) - e^{\frac{t\lambda}{2}}\lambda\sigma\right]^{n+2}} \tag{36}$$

for $0 \le t \le 2T$ and 0 otherwise, resulting in Equation (20).

Cherries-distance distribution

A cherry is a pair of adjacent tips on a tree (see Fig. 2). Let $N_\Lambda(t|T)dt$ be the average number of cherry pairs on the tree of length T, separated by the time distance t. Then, if the first branch splits at time T_1 and the two resulting subtrees possess M_1 and M_2 leaves, respectively, one gets the distribution in the form

$$N_\Lambda(t|T) = \left[2N_\Lambda(t|T-T_1) + P^2(1|T-T_1)\delta(t-2(T-T_1))I(0 \le t \le 2T)\right]e^{-\mu T_1}. \tag{37}$$

After average over T_1 and solving the resulting equation one obtains

$$N_\Lambda(t|T) = \frac{\lambda(\lambda-\mu)^4\sigma^2}{2}\frac{e^{\frac{t\lambda}{2}+T\lambda+\frac{3t\mu}{2}-T\mu}}{\left[e^{\frac{t\mu}{2}}(\mu-\lambda+\sigma\lambda) - e^{\frac{t\lambda}{2}}\lambda\sigma\right]^4} \tag{38}$$

for $0 \le t \le 2T$ and 0 otherwise, resulting in Equation (21).

The distribution of $N^1(t|T)dt$

In this Appendix we derive the distribution of $N^1(t|T)dt$. Consider an infinitesimal (in practice very small) interval, $[t, t+dt]$, such that $N(t|T)dt \ll 1$. The number of pairs $N^1(t|T)dt$ in this interval is distributed with the mean $N(t|T)dt$. The full distribution can be derived using the following arguments.

Pairs, separated by the time in the interval $[t, t+dt]$, branched at the time interval $[T-t/2-dt/2, T-t/2]$. The probability to have a branch in this interval is given by $\lambda e^{(\lambda-\mu)(T-t/2)}dt/2$. Given that there is a branching point in this interval it can lead to different number of leaves and, therefore, pairs separated by the time in the interval $[t, t+dt]$. The probability that no observed pairs survive from this branching is given by $1-[1-P(0|t/2)]^2$, where $P(n|T)$ is the probability to observe n leaves on a tree of age T and is given in Equations (3, 4). The probability that there are no observed pairs separated by the time in the interval $[t, t+dt]$ is given by Equation (25). The probability that there are $n > 0$ observed pairs separated by the time in the interval $[t, t+dt]$ is given by

$$\begin{aligned}\Pr\left(N^1(t|T)dt = n\right) &= \lambda e^{(\lambda-\mu)(T-t/2)}dt/2\sum_{n_1,n_2=1}^{n} P(n_1|t/2)P(n_2|t/2)\delta_{n_1 n_2,n} \\ &= \lambda e^{(\lambda-\mu)(T-t/2)}dt/2\sum_{n_1|n} P(n_1|t/2)P(n/n_1|t/2).\end{aligned} \tag{39}$$

The last sum runs over all divisors of n, including 1 and n. One can see the comparison of Equations (25) and (39) to numerical results in Fig. 7.

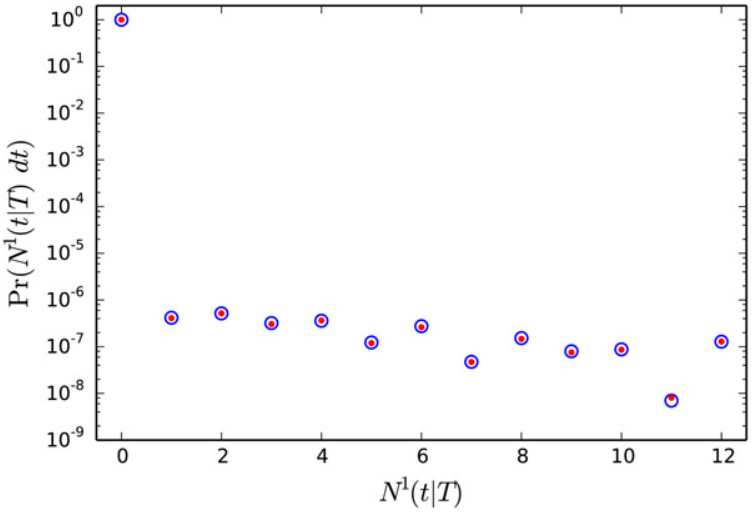

Fig 7. Probability to observe a certain number of pairs separated by the time in the interval [*t*, *t*+*dt*] on a tree of age *T*, $N^1(t|T)dt$. In this plot $T = 1$, $\lambda = 11$, $\mu = 5$, $\sigma = 0.01$, $t = 1.5$ and $dt = 0.00001$. Circles denote the results of numerical simulation and dots were obtained using the analytic formulas (25) for zero value and (39) for non-zero values. Note the gap between zero and non-zero probabilities due to small bin size, *dt*.

doi:10.1371/journal.pone.0120206.g007

Acknowledgments

The authors thank M. Mariadassou, P.W. Messer, and M. Vingron for helpful discussions.

Author Contributions

Conceived and designed the experiments: MS FM PA. Performed the experiments: MS FM PA. Wrote the paper: MS FM PA.

References

1. Yule G (1924) A mathematical theory of evolution, based on the conclusions of dr. jc willis. Philosophical Transactions of the Royal Society of London B B213: 21.
2. Karlin S, Taylor H (1975) A first course in stochastic processes. Academic Press, New York.
3. Newman ME (2005) Power laws, pareto distributions and zipf's law. Contemporary physics 46: 323. doi: 10.1080/00107510500052444
4. Novozhilov AS, Karev GP, Koonin EV (2006) Biological applications of the theory of birth-and-death processes. Briefings in bioinformatics 7: 70. doi: 10.1093/bib/bbk006 PMID: 16761366
5. Yanai I, Camacho CJ, DeLisi C (2000) Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Physical Review Letters 85: 2641. doi: 10.1103/PhysRevLett.85.2641 PMID: 10978127
6. Reed WJ, Hughes BD (2004) A model explaining the size distribution of gene and protein families. Mathematical biosciences 189: 97. doi: 10.1016/j.mbs.2003.11.002 PMID: 15051416
7. Tambovtsev Y, Martindale C (2007) Phoneme frequencies follow a yule distribution. SKASE Journal of Theoretical Linguistics 4: 1.
8. Raup DM (1985) Mathematical models of cladogenesis. Paleobiology 11: 42.
9. Aldous DJ (2001) Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. Statistical Science: 23.
10. Nee S, May RM, Harvey PH (1994) The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 344: 305. doi: 10.1098/rstb.1994.0068 PMID: 7938201

11. Nee S, Holmes EC, May RM, Harvey PH (1994) Extinction rates can be estimated from molecular phylogenies. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences 344: 77. doi: 10.1098/rstb.1994.0054 PMID: 8878259
12. Kendall DG (1949) Stochastic processes and population growth. Journal of the Royal Statistical Society Series B (Methodological) 11: 230.
13. Harvey PH, May RM, Nee S (1994) Phylogenies without fossils. Evolution: 523.
14. McKenzie A, Steel M (2000) Distributions of cherries for two models of trees. Mathematical biosciences 164: 81. doi: 10.1016/S0025-5564(99)00060-7 PMID: 10704639
15. Steel M, McKenzie A (2001) Properties of phylogenetic trees generated by yule-type speciation models. Mathematical biosciences 170: 91. doi: 10.1016/S0025-5564(00)00061-4 PMID: 11259805
16. Rosenberg NA (2006) The mean and variance of the numbers of r-pronged nodes and r-caterpillars in yule-generated genealogical trees. Annals of Combinatorics 10: 129. doi: 10.1007/s00026-006-0278-6
17. Mulder WH (2011) Probability distributions of ancestries and genealogical distances on stochastically generated rooted binary trees. Journal of theoretical biology 280: 139. doi: 10.1016/j.jtbi.2011.04.009 PMID: 21527261
18. Steel M, Mooers A (2010) The expected length of pendant and interior edges of a yule tree. Applied Mathematics Letters 23: 1315. doi: 10.1016/j.aml.2010.06.021
19. Mooers A, Gascuel O, Stadler T, Li H, Steel M (2012) Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. Systematic biology 61: 195. doi: 10.1093/sysbio/syr090 PMID: 21865336
20. Kumar S (2005) Molecular clocks: four decades of evolution. Nature Reviews Genetics 6: 654. doi: 10.1038/nrg1659 PMID: 16136655
21. Slatkin M, Hudson RR (1991) Pairwise comparisons of mitochondrial dna sequences in stable and exponentially growing populations. Genetics 129: 555. PMID: 1743491
22. Mora C, Tittensor DP, Adl S, Simpson AG, Worm B (2011) How many species are there on earth and in the ocean? PLoS biology 9: e1001127. doi: 10.1371/journal.pbio.1001127 PMID: 21886479
23. Pimm SL, Russell GJ, Gittleman JL, Brooks TM (1995) The future of biodiversity. Science: 347.
24. Hohna S, Stadler T, Ronquist F, Britton T (2011) Inferring speciation and extinction rates under different sampling schemes. Molecular biology and evolution 28: 2577. doi: 10.1093/molbev/msr095 PMID: 21482666
25. Kendall DG (1948) On some modes of population growth leading to ra fisher's logarithmic series distribution. Biometrika: 6.
26. Massip F, Sheinman M, Schbath S and Arndt PF (2015) How Evolution of Genomes Is Reflected in Exact DNA Sequence Match Statistic. Mol. Biol. Evol. 32(2): 524. doi: 10.1093/molbev/msu313 PMID: 25398628
27. Stadler T (2009) On incomplete sampling under birth–death models and connections to the sampling-based coalescent. Journal of Theoretical Biology 261: 58. doi: 10.1016/j.jtbi.2009.07.018 PMID: 19631666
28. Jetz W, Thomas G, Joy J, Hartmann K, Mooers A (2012) The global diversity of birds in space and time. Nature 491: 444. doi: 10.1038/nature11631 PMID: 23123857
29. Holland PW, Welsch RE (1977) Robust regression using iteratively reweighted least-squares. Communications in Statistics-Theory and Methods 6: 813. doi: 10.1080/03610927708827533
30. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, et al. (2014) Beast 2: a software platform for bayesian evolutionary analysis. PLoS computational biology 10: e1003537. doi: 10.1371/journal.pcbi.1003537 PMID: 24722319
31. Ronquist F, Huelsenbeck JP (2003) Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572. doi: 10.1093/bioinformatics/btg180 PMID: 12912839

Bibliography

[1] E. Chargaff, "Chemical specificity of nucleic acids and mechanism of their enzymatic degradation," Cellular and Molecular Life Sciences **6**, 201 (1950).

[2] R. E. Franklin and R. G. Gosling, "Molecular configuration in sodium thymonucleate," Nature **171**, 740 (1953).

[3] J. D. Watson, F. H. Crick, et al., "Molecular structure of nucleic acids," Nature **171**, 737 (1953).

[4] P. Yakovchuk, E. Protozanova, and M. D. Frank-Kamenetskii, "Base-stacking and base-pairing contributions into thermal stability of the DNA double helix," Nucleic Acids Research **34**, 564 (2006).

[5] M. Bansal, "DNA structure: Revisiting the watson-crick double helix," Current Science **85**, 1556 (2003).

[6] E. P. Consortium et al., "An integrated encyclopedia of dna elements in the human genome," Nature **489**, 57 (2012).

[7] D. Graur, Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik, "On the immortality of television sets:"function" in the human genome according to the evolution-free gospel of encode," Genome Biology and Evolution **5**, 578 (2013).

[8] M. Kellis, B. Wold, M. P. Snyder, B. E. Bernstein, A. Kundaje, G. K. Marinov, L. D. Ward, E. Birney, G. E. Crawford, J. Dekker, et al., "Defining functional dna elements in the human genome," Proceedings of the National Academy of Sciences **111**, 6131 (2014).

[9] B. Prum, F. Rodolphe, and É. de Turckheim, "Finding words with unexpected frequencies in deoxyribonucleic acid sequences," Journal of the Royal Statistical Society. Series B (Methodological) pp. 205–220 (1995).

[10] S. Schbath, B. Prum, and E. de Turckheim, "Exceptional motifs in different markov chain models for a statistical analysis of dna sequences," Journal of Computational Biology **2**, 417 (1995).

[11] P. A. Pevzner, M. Y. Borodovsky, and A. A. Mironov, "Linguistics of nucleotide sequences i: the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words," Journal of Biomolecular Structure and Dynamics **6**, 1013 (1989).

[12] S. Robin, F. Rodolphe, and S. Schbath, *DNA, words and models: Statistics of exceptional words* (Cambridge University Press, 2005).

[13] S. Karlin and H. M. Taylor, "A first course in stochastic processes," Academic Press, New York (1975).

[14] C. Burge and S. Karlin, "Prediction of complete gene structures in human genomic dna," Journal of Molecular Biology **268**, 78 (1997).

[15] T. H. Jukes and C. R. Cantor, "Evolution of protein molecules," Mammalian Protein Metabolism **3**, 21 (1969).

[16] R. Rudner, J. D. Karkas, and E. Chargaff, "Separation of b. subtilis DNA into complementary strands, i. biological properties.," Proceedings of the National Academy of Sciences **60**, 630 (1968).

[17] C. Nikolaou and Y. Almirantis, "Deviations from chargaff's second parity rule in organellar DNA: Insights into the evolution of organellar genomes," Gene **381**, 34 (2006).

[18] D. Mitchell and R. Bridge, "A test of chargaff's second rule," Biochemical and Biophysical Research Communications **340**, 90 (2006).

[19] J. Lobry and C. Lobry, "Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant.," Molecular Biology and Evolution **16**, 719 (1999).

[20] N. Sueoka, "Intrastrand parity rules of DNA base composition and usage biases of synonymous codons," Journal of Molecular Evolution **40**, 318 (1995).

[21] S. Aerts, G. Thijs, M. Dabrowski, Y. Moreau, and B. De Moor, "Comprehensive analysis of the base composition around the transcription start site in metazoa," BMC Genomics **5**, 34 (2004).

[22] P. Polak and P. F. Arndt, "Transcription induces strand-specific mutations at the 5' end of human genes," Genome Research **18**, 1216 (2008).

[23] P. Polak and P. F. Arndt, "Long-range bidirectional strand asymmetries originate at cpg islands in the human genome," Genome Biology and Evolution **1**, 189 (2009).

[24] M. Touchon, A. Arneodo, Y. d'Aubenton Carafa, and C. Thermes, "Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes," Nucleic Acids Research **32**, 4969 (2004).

[25] M. Touchon and E. P. Rocha, "From GC skews to wavelets: a gentle guide to the analysis of compositional asymmetries in genomic data," Biochimie **90**, 648 (2008).

[26] N. V. Sernova and M. S. Gelfand, "Identification of replication origins in prokaryotic genomes," Briefings in Bioinformatics **9**, 376 (2008).

[27] S. Karlin and S. F. Altschul, "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes," Proceedings of the National Academy of Sciences **87**, 2264 (1990).

[28] S. Karlin and S. F. Altschul, "Applications and statistics for multiple high-scoring segments in molecular sequences," Proceedings of the National Academy of Sciences **90**, 5873 (1993).

[29] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al., "Initial sequencing and analysis of the human genome," Nature **409**, 860 (2001).

[30] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, et al., "The b73 maize genome: complexity, diversity, and dynamics," Science **326**, 1112 (2009).

[31] A. G. Initiative et al., "Analysis of the genome sequence of the flowering plant arabidopsis thaliana.," Nature **408**, 796 (2000).

[32] J. Jurka and T. Smith, "A fundamental division in the alu family of repeated sequences," Proceedings of the National Academy of Sciences **85**, 4775 (1988).

[33] R. J. Britten, W. F. Baron, D. B. Stout, and E. H. Davidson, "Sources and evolution of human alu repeated sequences," Proceedings of the National Academy of Sciences **85**, 4770 (1988).

[34] M. A. Batzer and P. L. Deininger, "Alu repeats and human genomic diversity," Nature Reviews Genetics **3**, 370 (2002).

[35] P. L. Deininger and M. A. Batzer, "Mammalian retroelements," Genome research **12**, 1455 (2002).

[36] R. Cordaux and M. A. Batzer, "The impact of retrotransposons on human genome evolution," Nature Reviews Genetics **10**, 691 (2009).

[37] B. Chénais, A. Caruso, S. Hiard, and N. Casse, "The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments," Gene **509**, 7 (2012).

[38] I. K. Jordan, I. B. Rogozin, G. V. Glazko, and E. V. Koonin, "Origin of a substantial fraction of human regulatory sequences from transposable elements," Trends in Genetics **19**, 68 (2003).

[39] K. R. Oliver, J. A. McComb, and W. K. Greene, "Transposable elements: powerful contributors to angiosperm evolution and diversity," Genome Biology and Evolution **5**, 1886 (2013).

[40] J. K. Pace and C. Feschotte, "The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage," Genome Research **17**, 422 (2007).

[41] S. Ohno, "Evolution by gene duplication,", Springer (1970).

[42] M. Lynch and J. S. Conery, "The evolutionary fate and consequences of duplicate genes," Science **290**, 1151 (2000).

[43] H. Innan and F. Kondrashov, "The evolution of gene duplications: classifying and distinguishing between models," Nature Reviews Genetics **11**, 97 (2010).

[44] J. A. Bailey and E. E. Eichler, "Primate segmental duplications: crucibles of evolution, diversity and disease," Nature Reviews Genetics **7**, 552 (2006).

[45] X. She, Z. Cheng, S. Zöllner, D. M. Church, and E. E. Eichler, "Mouse segmental duplication and copy number variation," Nature Genetics **40**, 909 (2008).

[46] J. A. Bailey, G. Liu, and E. E. Eichler, "An alu transposition model for the origin and expansion of human segmental duplications," The American Journal of Human Genetics **73**, 823 (2003).

[47] P. M. Kim, H. Y. Lam, A. E. Urban, J. O. Korbel, J. Affourtit, F. Grubert, X. Chen, S. Weissman, M. Snyder, and M. B. Gerstein, "Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history," Genome Research **18**, 1865 (2008).

[48] J. K. Moore and J. E. Haber, "Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in saccharomyces cerevisiae.," Molecular and Cellular Biology **16**, 2164 (1996).

[49] E. V. Linardopoulou, E. M. Williams, Y. Fan, C. Friedman, J. M. Young, and B. J. Trask, "Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication," Nature **437**, 94 (2005).

[50] P. Hastings, J. R. Lupski, S. M. Rosenberg, and G. Ira, "Mechanisms of change in gene copy number," Nature Reviews Genetics **10**, 551 (2009).

[51] X. She, J. E. Horvath, Z. Jiang, G. Liu, T. S. Furey, L. Christ, R. Clark, T. Graves, C. L. Gulden, C. Alkan, et al., "The structure and evolution of centromeric transition regions within the human genome," Nature **430**, 857 (2004).

[52] E. F. Vanin, "Processed pseudogenes: characteristics and evolution," Annual Review of Genetics **19**, 253 (1985).

[53] H. Kaessmann, N. Vinckenbosch, and M. Long, "Rna-based gene duplication: mechanistic and evolutionary insights," Nature Reviews Genetics **10**, 19 (2009).

[54] K. Okamura and K. Nakai, "Retrotransposition as a source of new promoters," Molecular Biology and Evolution **25**, 1231 (2008).

[55] O. Podlaha and J. Zhang, "Processed pseudogenes: the 'fossilized footprints' of past gene expression," Trends in Genetics **25**, 429 (2009).

[56] L. McDonell, G. Drouin, and T. Bureau, "The abundance of processed pseudogenes derived from glycolytic genes is correlated with their expression level," Genome **55**, 147 (2012).

[57] O. Jaillon, J.-M. Aury, and P. Wincker, ""changing by doubling", the impact of whole genome duplications in the evolution of eukaryotes," Comptes Rendus Biologies **332**, 241 (2009).

[58] T. E. Wood, N. Takebayashi, M. S. Barker, I. Mayrose, P. B. Greenspoon, and L. H. Rieseberg, "The frequency of polyploid speciation in vascular plants," Proceedings of the National Academy of Sciences **106**, 13875 (2009).

[59] B. C. Thomas, B. Pedersen, and M. Freeling, "Following tetraploidy in an arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes," Genome Research **16**, 934 (2006).

[60] F. Massip and P. F. Arndt, "Neutral evolution of duplicated dna: an evolutionary stick-breaking process causes scale-invariant behavior," Physical Review Letters **110**, 148101 (2013).

[61] F. Massip, M. Sheinman, S. Schbath, and P. F. Arndt, "How evolution of genomes is reflected in exact DNA sequence match statistics," Molecular Biology and Evolution **32**, 524 (2015).

[62] M. Sheinman, F. Massip, and P. F. Arndt, "Statistical properties of pairwise distances between leaves on a random yule tree," PLoS ONE **10**, e0120206 (2015).

[63] J. A. Lee, C. M. Carvalho, and J. R. Lupski, "A dna replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders," Cell **131**, 1235 (2007).

[64] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, "Versatile and open software for comparing large genomes," Genome Biology **5**, R12 (2004).

[65] D. DeSolla Price, "Networks of scientific papers," Science pp. 510–515 (1965).

[66] M. E. Newman, "Power laws, pareto distributions and zipf's law," Contemporary Physics **46**, 323 (2005).

[67] N. M. Luscombe, J. Qian, Z. Zhang, T. Johnson, and M. Gerstein, "The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties," Genome Biology **3**, 1 (2002).

[68] S. S. Sindi, B. R. Hunt, and J. A. Yorke, "Duplication count distributions in DNA sequences," Physical Review E **78**, 061912 (2008).

[69] C. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. Stanley, et al., "Long-range correlations in nucleotide sequences," Nature **356**, 168 (1992).

[70] W. J. Reed and B. D. Hughes, "From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature," Physical Review E **66**, 067103 (2002).

[71] T. Gisiger, "Scale invariance in biology: coincidence or footprint of a universal mechanism?," Biological Reviews **76**, 161 (2007).

[72] E. van Nimwegen, in *Power Laws, Scale-Free Networks and Genome Biology* (Springer, 2006), pp. 236–253.

[73] M. P. H. Stumpf and M. A. Porter, "Mathematics. Critical truths about power laws.," Science **335**, 665 (2012).

[74] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," SIAM Review **51**, 661 (2009).

[75] P. W. Messer, P. F. Arndt, and M. Lässig, "Solvable sequence evolution models and genomic correlations," Physical Review Letters **94**, 138103 (2005).

[76] G. Lima-Mendez and J. van Helden, "The powerful law of the power law and other myths in network biology," Molecular BioSystems **5**, 1482 (2009).

[77] D. Sellis, A. Provata, and Y. Almirantis, "Alu and line1 distributions in the human chromosomes: evidence of global genomic organization expressed in the form of power laws," Molecular biology and evolution **24**, 2385 (2007).

[78] M. L. Goldstein, S. A. Morris, and G. G. Yen, "Problems with fitting to the power-law distribution," The European Physical Journal B-Condensed Matter and Complex Systems **41**, 255 (2004).

[79] S. Milojević, "Power law distributions in information science: Making the case for logarithmic binning," Journal of the American Society for Information Science and Technology **61**, 2417 (2010).

[80] B. M. Hill et al., "A simple general approach to inference about the tail of a distribution," The annals of statistics **3**, 1163 (1975).

[81] G. U. Yule, "A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs," Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character pp. 21–87 (1925).

[82] D. G. Kendall, "Stochastic processes and population growth," Journal of the Royal Statistical Society. Series B (Methodological) **11**, 230 (1949).

[83] F. Cunningham, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, et al., "Ensembl 2015," Nucleic Acids Research **43**, D662 (2015).

[84] P. Lamesch, T. Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, R. Muller, K. Dreher, D. L. Alexander, M. Garcia-Hernandez, et al., "The arabidopsis information resource (tair): improved gene annotation and new tools," Nucleic Acids Research **40**, D1202 (2012).

[85] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," Nucleic Acids Research **25**, 3389 (1997).

[86] K. Katoh and D. M. Standley, "Mafft multiple sequence alignment software version 7: improvements in performance and usability," Molecular Biology and Evolution **30**, 772 (2013).

[87] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, "trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses," Bioinformatics **25**, 1972 (2009).

[88] J. Felsenstein, "Phylip - phylogeny inference package (version 3.2)," Cladistics **5**, 164 (1989).

[89] C. G. Felsenstein J, "A hidden markov model approach to variation among sites in rate of evolution, and the branching order in hominoidea," Molecular Biology and Evolution **13**, 93 (1996).

[90] A. Smit, R. Hubley, and P. Green, "Repeatmasker open-3.0," See http://www. repeatmasker. org (1996–2010).

[91] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences.," Journal of Molecular Biology **147**, 195 (1981).

[92] O. Gotoh, "An improved algorithm for matching biological sequences," Journal of Molecular Biology **162**, 705 (1982).

[93] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz, "Repbase Update, a database of eukaryotic repetitive elements.," Cytogenetic and Genome Research **110**, 462 (2005).

[94] K. Gao and J. Miller, "Algebraic distribution of segmental duplication lengths in whole-genome sequence self-alignments," PLoS ONE **6**, e18464 (2011).

[95] S. S. Sindi, Ph.D. thesis (2006).

[96] M. Csűrös, L. Noé, and G. Kucherov, "Reconsidering the significance of genomic word frequencies," Trends in Genetics **23**, 543 (2007).

[97] W. Salerno, P. Havlak, and J. Miller, "Scale-invariant structure of strongly conserved sequence in genomic intersections and alignments," Proceedings of the National Academy of Sciences **103**, 13121 (2006).

[98] W. Kuhn, "Über die Kinetik des Abbaues hochmolekularer Ketten," Berichte der Deutschen Chemischen Gesellschaft **63**, 1502 (1930).

[99] R. M. Ziff and E. D. McGrady, "The kinetics of cluster fragmentation and depolymerisation," Journal of Physics A: Mathematical and General **18**, 3027 (1985).

[100] W. Reed and B. Hughes, "From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature," Physical Review E **66**, 067103 (2002).

[101] E. Ben-Naim and P. Krapivsky, "Fragmentation with a steady source," Physics Letters A **275**, 48 (2000).

[102] R. V. Samonte and E. E. Eichler, "Segmental duplications and the evolution of the primate genome," Nature Reviews Genetics **3**, 65 (2002).

[103] D. F. Conrad, J. E. Keebler, M. A. DePristo, S. J. Lindsay, Y. Zhang, F. Casals, Y. Idaghdour, C. L. Hartl, C. Torroja, K. V. Garimella, et al., "Variation in genome-wide mutation rates within and between human families," Nature **201** (2011).

[104] D. C. Phillips, *The three-dimensional structure of an enzyme molecule* (WH Freeman and Company, 1966).

[105] L. Oddershede, P. Dimon, and J. Bohr, "Self-organized criticality in fragmenting.," Physical Review Letters **71**, 3107 (1993).

[106] E. Taillefer and J. Miller, "Exhaustive computation of exact duplications via super and non-nested local maximal repeats," Journal of Bioinformatics and Computational Biology **12** (2014).

[107] L. Zhang, H. H. Lu, W.-y. Chung, J. Yang, and W.-H. Li, "Patterns of segmental duplication in the human genome," Molecular Biology and Evolution **22**, 135 (2005).

[108] T. Marques-Bonet, J. M. Kidd, M. Ventura, T. A. Graves, Z. Cheng, L. W. Hillier, Z. Jiang, C. Baker, R. Malfavon-Borja, L. A. Fulton, et al., "A burst of segmental duplications in the genome of the african great ape ancestor," Nature **457**, 877 (2009).

[109] J. E. Karro, Y. Yan, D. Zheng, Z. Zhang, N. Carriero, P. Cayting, P. Harrison, and M. Gerstein, "Pseudogene. org: a comprehensive database and comparison platform for pseudogene annotation," Nucleic Acids Research **35**, D55 (2007).

[110] Y. Nakatani, H. Takeda, Y. Kohara, and S. Morishita, "Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates," Genome Research **17**, 1254 (2007).

[111] Y. Van de Peer, "Computational approaches to unveiling ancient genome duplications," Nature Reviews Genetics **5**, 752 (2004).

[112] Q. Zhang and N. Backström, "Assembly errors cause false tandem duplicate regions in the chicken (gallus gallus) genome sequence," Chromosoma **123**, 165 (2014).

[113] D. P. Locke, L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S.-P. Yang, Z. Wang, A. T. Chinwalla, P. Minx, et al., "Comparative and demographic analysis of orang-utan genomes," Nature **469**, 529 (2011).

[114] K. Han, M. K. Konkel, J. Xing, H. Wang, J. Lee, T. J. Meyer, C. T. Huang, E. Sandifer, K. Hebert, E. W. Barnes, et al., "Mobile DNA in old world monkeys: a glimpse through the rhesus macaque genome," Science **316**, 238 (2007).

[115] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg, "Versatile and open software for comparing large genomes.," Genome Biology **5**, R12 (2004).

[116] K. Gao and J. Miller, "Human-chimpanzee alignment: Ortholog exponentials and paralog power laws," Computational Biology and Chemistry (2014).

[117] I. Gonçalves, L. Duret, and D. Mouchiroud, "Nature and structure of human genes that generate retropseudogenes," Genome Research **10**, 672 (2000).

[118] S. Balasubramanian, D. Zheng, Y.-J. Liu, G. Fang, A. Frankish, N. Carriero, R. Robilotto, P. Cayting, and M. Gerstein, "Comparative analysis of processed ribosomal protein pseudogenes in four mammalian genomes," Genome Biology **10**, R2 (2009).

[119] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, and D. Haussler, "Ultraconserved elements in the human genome," Science **304**, 1321 (2004).

[120] E. T. Dermitzakis, A. Reymond, and S. E. Antonarakis, "Conserved non-genic sequences–an unexpected feature of mammalian genomes," Nature Reviews Genetics **6**, 151 (2005).

[121] L. Duret, "Neutral theory: the null hypothesis of molecular evolution," Nature Education **1**, 218 (2008).

[122] H. Flanders, "Differentiation under the integral sign," The American Mathematical Monthly **80**, pp. 615 (1973), ISSN 00029890.

[123] M. Nei and S. Kumar, *Molecular evolution and phylogenetics* (Oxford University Press, 2000).

[124] C.-I. Wu and W.-H. Li, "Evidence for higher rates of nucleotide substitution in rodents than in man," Proceedings of the National Academy of Sciences **82**, 1741 (1985).

[125] C. Simillion, K. Vandepoele, M. C. Van Montagu, M. Zabeau, and Y. Van de Peer, "The hidden duplication past of arabidopsis thaliana," Proceedings of the National Academy of Sciences **99**, 13627 (2002).

[126] A. Meyer and Y. Van de Peer, "From 2r to 3r: evidence for a fish-specific genome duplication (fsgd)," Bioessays **27**, 937 (2005).

[127] M. Kasahara, "The 2R hypothesis: an update," Current Opinion in Immunology **19**, 547 (2007).

[128] I. Mayrose, S. H. Zhan, C. J. Rothfels, K. Magnuson-Ford, M. S. Barker, L. H. Rieseberg, and S. P. Otto, "Recently formed polyploid plants diversify at lower rates," Science **333**, 1257 (2011).

[129] M. R. Woodhouse, J. C. Schnable, B. S. Pedersen, E. Lyons, D. Lisch, S. Subramaniam, and M. Freeling, "Following tetraploidy in maize, a short deletion

mechanism removed genes preferentially from one of the two homeologs," PLoS Biology **8**, e1000409 (2010).

[130] S. Ohno, in *Seminars in cell & developmental biology* (Elsevier, 1999), vol. 10, pp. 517–522.

[131] A. McLysaght, K. Hokamp, and K. H. Wolfe, "Extensive genomic duplication during early chordate evolution," Nature Genetics **31**, 200 (2002).

[132] P. Dehal and J. L. Boore, "Two rounds of whole genome duplication in the ancestral vertebrate," PLoS Biology **3**, e314 (2005).

[133] Y. Nakatani, H. Takeda, Y. Kohara, and S. Morishita, "Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates," Genome Research **17**, 1254 (2007).

[134] C. L. McGrath, J.-F. Gout, P. Johri, T. G. Doak, and M. Lynch, "Differential retention and divergent resolution of duplicate genes following whole-genome duplication," Genome Research **24**, 1665 (2014).

[135] T. Makino and A. McLysaght, "Positionally biased gene loss after whole genome duplication: evidence from human, yeast, and plant," Genome Research **22**, 2427 (2012).

[136] L. Duret and N. Galtier, "Biased gene conversion and the evolution of mammalian genomic landscapes," Annual Review of Genomics and Human Genetics **10**, 285 (2009).

[137] H. Jellinek and G. White, "The degradation of long-chain molecules by ultrasonic waves. II. Degradation of polystyrene," Journal of Polymer Science **6**, 757 (1951).

Titre : Le devenir statistique de l'ADN génomique : Modélisation des statistiques d'appariement dans différents scénarios évolutifs

Mots clés : Duplications, Distributions en loi Puissance, Modèles évolutifs, Propriétés statistiques des Génomes

Le but de cette thèse est d'étudier la distribution des tailles des répétitions au sein d'un même génome, ainsi que la distribution des tailles des appariements obtenus en comparant différents génomes. Ces distributions présentent d'importantes déviations par rapport aux prédictions des modèles probabilistes existants. Étonnamment, les déviations observées sont distribuées selon une loi de puissance. Afin d'étudier ce phénomène, nous avons développé des modèles mathématiques prenant en compte des mécanismes évolutifs plus complexes, et qui expliquent les distributions observées. Nous avons aussi implémenté des modèles d'évolution de séquences *in silico* générant des séquences ayant les mêmes propriétés que les génomes étudiés. Enfin, nous avons montré que nos modèles permettent de tester la qualité des génomes récemment séquencés, et de mettre en évidence la prévalence de certains mécanismes évolutifs dans les génomes eucaryotes.

Title : The Statistical Fate of Genomic DNA :Modelling Match Statistics in Different Evolutionary Scenarios

Keywords : Duplications, Scale Free Distributions, Evolutionary Models, Statistical Properties of Genomes

Abstract : In this thesis, we study the length distribution of maximal exact matches within and between eukaryotic genomes. These distributions strongly deviate from what one could expect from simple probabilistic models and, surprisingly, present a power-law behavior. To analyze these deviations, we develop mathematical frameworks taking into account complex mechanisms and that reproduce the observed deviations. We also implemented *in silico* sequence evolution models that reproduce these behaviors. Finally, we show that we can use our framework to assess the quality of sequences of recently sequenced genomes and to highlight the importance of unexpected biological mechanisms in eukaryotic genomes.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France