Models and Algorithms for Whole-Genome Evolution and their Use in Phylogenetic Inference

THÈSE N^{O} 5435 (2012)

PRÉSENTÉE LE 31 JUILLET 2012
À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE DE BIOLOGIE COMPUTATIONNELLE ET BIOINFORMATIQUE
PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Yu LIN

acceptée sur proposition du jury:

Prof. R. Guerraoui, président du jury
Prof. B. Moret, directeur de thèse
Prof. J. Maddocks, rapporteur
Prof. C. Papadimitriou, rapporteur
Prof. R. Schwartz, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2012

Abstract

The rapid accumulation of sequenced genomes offers the chance to resolve longstanding questions about the evolutionary histories, or phylogenies, of groups of organisms. The relatively rare occurrence of large-scale evolutionary events in a whole genome, events such as genome rearrangements, duplications and losses, enables us to extract a strong and robust phylogenetic signal from whole-genome data. The work presented in this dissertation focuses on models and algorithms for whole-genome evolution and their use in phylogenetic inference. We designed algorithms to estimate pairwise genomic distances from large-scale genomic changes. We refined the evolutionary models on whole-genome evolution. We also made use of these results to provide fast and accurate methods for phylogenetic inference, that scales up, in both speed and accuracy, to modern high-resolution whole-genome data.

We designed algorithms to estimate the true evolutionary distance between two genomes under genome rearrangements, and also under rearrangements, plus gains and losses. We refined the evolutionary model to be the first mathematical model to preserve the structural dichotomy in genomic organization between most prokaryotes and most eukaryotes. Those models and associated distance estimators provide a basis for studying facets of possible mechanisms of evolution through simulation and application to real genomes.

Phylogenetic analyses from whole-genome data have been limited to small collections of genomes and low-resolution data; they have also lacked an effective assessment of robustness. We developed an approach that combines our distance estimator, any standard distance-based reconstruction algorithm, and a novel bootstrapping method based on resampling genomic adjacencies. The resulting tool overcomes a serious and longstanding impediment to the use of whole-genome data in phylogenetic inference and provides results comparable in accuracy and robustness to distance-based methods for sequence data.

Maximum-likelihood approaches have been successfully applied to phylogenetic inferences for aligned sequences, but such applications remain primitive for whole-genome data. We developed a maximum-likelihood approach to phylogenetic analysis from whole-genome data. In combination with our bootstrap scheme, this new approach yields the first reliable phylogenetic tool for the analysis of whole-genome data at the level of syntenic blocks.

keywords: genome rearrangement, distance estimation, distance-based reconstruction, bootstrap, maximum-likelihood

Résumé

L’accumulation rapide de génomes séquencés offre l’opportunité de résoudre des questions de longue date sur l’évolution, en particulier sur les phylogénies. Cette étude est possible grâce au faible nombre d’événements génomiques à large échelle, évńements tels que les réarrangements génomiques, ou encore les duplications et pertes de segments. Le travail présenté ici examine les modèles et algorithmes pour l’évolution de génomes entiers et leur utilisation dans l’inférence phylogénétique. Nous avons conçu des algorithmes pour estimer la distance génomique par paires à partir de changements génomiques à large échelle et affiné les modèles évolutifs associés. Nous avons aussi utilisé ces résultats pour fournir une méthode d’inférence phylogénétique précise et rapide et avons conçu des approches pour le calcul des scores de bootstrap des arbres résultants.

Nous avons également conçu des algorithmes pour estimer la véritable distance évolutive entre deux génomes soumis à des réarrangements génomiques et, dans un deuxième cas, incluant des gains et pertes de segments. Ces estimateurs décrivent le comportement asymptotique de la structure d’un génome et peuvent être utilisés pour la prédiction. Notre modèle affiné est le premier modèle mathématique à préserver la dichotomie structurale dans l’organisation génomique entre la plupart des procaryotes et les eucaryotes. Ces modèles et leurs estimateurs de distance associés fournissent une base pour l’étude des différentes facettes des mécanismes d’évolution possibles.

Les analyses phylogénétiques pour les génomes entiers ont été limitées à des petites collections de génomes contenants des données de basse résolution; de plus, une appréciation efficace de leur robustesse a manqué. Avec l’utilisation de notre estimateur de distance décrit ci-dessus, nous avons développé une approche pour l’inférence phylogénétique suffisament rapide et précise pour utiliser à plein les nouvelles données de haute résolution pour les génomes entiers. Cette approche combine notre estimateur de distance avec n’importe quelle méthode pour la reconstruction phylogénétique basée sur la distance et une nouvelle méthode de bootstrapping basée sur un ré-échantillonnage de contiguïtés génomiques. L’outil résultant surmonte un sérieux et ancien défaut inhérent à l’usage de données de génomes entier pour l’inférence phylogénétique et fournit des résultats comparables en précision et robustesse aux méthodes basées sur la distance pour les données de séquences.

Les approches basées sur le maximum de vraisemblance ont été appliquées avec succès pour l’inférence phylogénétique à base de séquences alignées mais restent primitives pour les données de génomes entiers. Nous avons développé une méthode de maximum de vraisemblance pour les données de génomes entiers en utilisant notre estimateur pour calculer les probabilités de transition et en encodant l’information des contiguïtés et du contenu génomique sur des séquences binaires. En combinaison avec notre nouveau schéma de bootstrap, cette approche produit le premier outil phylogénétique pour l’analyse de génomes entiers au niveau de blocs synténiques.

mots clefs: réarrangements génomiques, estimation de distance, reconstruction basée sur la distance, bootstrap, maximum de vraisemblance

Acknowledgments

Bernard Moret is the best role model for a scientist and teacher. It has been my privilege to have him as my PhD advisor. His great mind and personal charisma filled me with admiration. He granted me the freedom to discover my real passion in research, and also provided continuous support, guidance and encouragement along the way. Thank you, Bernard!

During my time in Bernard's lab, I have been blessed with many friendly and cheerful colleagues. I would like to express my gratefulness to Krister Swenson for introducing me to this research area and guiding my research. I was fortunate to have befriended and collaborated closely with Vaibhav Rajan, with whom I spent a pleasant and memorable time in working and travelling. I am grateful to Xiuwei Zhang for recommending this lab to me and being supportive all the time. I appreciate Yann Christinat's help, particularly in teaching me skiing and answering my French questions. I also would like to extend my sincere gratitude to other members in our lab: Cristina Ghiurcuta, Nishanth Nair, Mingfu Shao and Min Ye, who gave me company as my academic sisters and brothers.

I am deeply indebted to Prof. Rachid Guerraoui, Prof. John Maddocks, Prof. Russell Schwartz and Prof. Christos Papadimitriou for their willingness to serve on my thesis committee. It has been a great pleasure and honor for me to receive valuable insights and detailed comments from Prof. Christos Papadimitriou.

I would like to acknowledge my family and friends for their continued love and support throughout all my endeavors.

Finally, my special thanks goes to Dongbo Bu and Bernard Moret, who patiently and skillfully encouraged me to pursue an academic career.

Contents

1 Introduction 9

1.1 Data and models of molecular evolution . . . 9

1.1.1 Sequence data . . . 9

1.1.2 Whole-genome data (at the level of syntenic blocks) . . . 10

1.2 Methods for phylogenetic reconstruction . . . 10

1.2.1 Distance-based methods . . . 10

1.2.2 Parsimony-based methods . . . 11

1.2.3 Likelihood-based methods . . . 11

1.3 Handling whole-genome data . . . 11

1.4 Contributions in this dissertation . . . 12

1.4.1 Models and distance estimation on whole-genome evolution . . . 12

1.4.2 Distance-based reconstruction with bootstrapping from whole-genome data . 13

1.4.3 Maximum-likelihood reconstruction from whole-genome data . . . 13

2 Models and distance estimation on whole-genome evolution 15

2.1 Estimating true evolutionary distances under the DCJ model . . . 16

2.1.1 Preliminaries on whole-genome data and the DCJ model . . . 16

2.1.2 True distance estimation under the DCJ model . . . 18

2.1.3 Experimental results . . . 21

2.1.4 Discussion . . . 24

2.2 Models and distance estimations under rearrangements, duplications, and losses . . . 24

2.2.1 Preliminaries on gene-order data and the new evolutionary model . . . 25

2.2.2 True distance estimation under the new evolutionary model . . . 28

2.2.3 Model characteristics . . . 32

2.2.4 Experimental results . . . 35

2.2.5 Discussion . . . 38

3 Distance-based reconstruction with bootstrapping from whole-genome data 41

3.1 Distance-based reconstruction from whole-genome data . . . 42

3.1.1 Phylogenetic reconstruction and accuracy testing . . . 42

3.1.2 Experimental design . . . 42

3.1.3 Experimental results . . . 43

3.1.4 Discussion . . . 45

3.2 Bootstrapping phylogenies . . . 45

3.2.1 Robustness estimation for trees reconstructed from whole-genome data . . . 47

3.2.2 Experimental design . . . 48

3.2.3 Experimental results . . . 49

3.2.4 Discussion . . . 50

4 Maximum-likelihood reconstruction from whole-genome data 55
4.1 Methods 55
4.2 Experimental design 57
4.3 Experimental results 58
4.4 Discussion 63

5 Conclusion and discussion 65

Chapter 1

Introduction

One of the most exciting aspects of our planet is the diversity of life. In spite of 250 years of taxonomic classification and over 1.2 million species already catalogued in a central database, millions of species on Earth still await description [72].

All life arises by evolution. One central goal in the study of evolution is to infer the evolutionary history, or phylogeny, that unites all life on Earth. Phylogenetics is central to biological studies. For example, phylogenies provide new ways to represent and measure the diversity of life, and retain the information that can help us preserve the pattern of biodiversity as well as the processes that have generated the pattern [68, 112]. Comparative studies make extensive use of phylogenies in making predictions about species and their biogeographical, ecological, physiological, behavioral, developmental or genomic significance. This predictive power has in turn proven useful in practical areas such as prospecting for novel chemicals and medicines, developing control measure for pests, guiding biotechnology, and evaluating potential cures for diseases. Phylogenies are invaluable to the study of modern epidemiology, e.g., in identifying and classifying emerging viruses such as SARS [25], and in understanding the genetic evolution of HIV [89] and influenza [33].

Evolution takes place over long periods of time, and thus cannot be observed directly. In reconstructing phylogenies, the main challenge is that we lack the information on evolutionary events that occurred in the past. Although fossil records contain morphological characters of ancient species, they are often difficult to interpret and incorporate into phylogenies. Thus we rely on contemporary data and a model of evolution to understand the past, and design reconstruction methods to infer the phylogeny. While many types of data are available, the dominant choice today is molecular data. Molecular data has the significant advantage of being exact, reproducible and easy to obtain. Further, each nucleotide in a DNA or RNA sequence is, by itself, a well defined discrete character. While genomic sequences remains the main source of molecular data, promising new types of genomic data are appearing, most notably whole-genome data [76].

In the following sections of this chapter we will provide a brief background on the data, models and methods used in phylogenetic reconstruction and provide the context for the specific problems we have addressed in this dissertation.

1.1 Data and models of molecular evolution

1.1.1 Sequence data

In sequence data, characters, individual positions in the sequence, can be assumed to be in one of a few states, e.g., 4 states for nucleotides or 20 states for amino acids. Such data evolve through evolutionary events such as point mutations, insertions and deletions. Pioneering work on models of evolution on sequence data began during the 1960s. For example, Zuckerkandl and Pauling proposed the molecular clock theory: they suggested that the rate of evolutionary change of any specified protein was approximately constant over time and over different lineages [123]. Jukes and Cantor proposed a stochastic model for DNA substitution, assuming equal transition rates as well as equal equilibrium frequencies for all bases [54]. Kimura introduced a model that distinguishes two types of substitutions, transitions and transversions, while still assuming equal base frequencies [55]. F84 (in the PHYLIP package) and HKY models [41] are two widely used models that allow arbitrary base frequencies [29]. More recently, many sophisticated models were developed and refined to account for the inherent complexity of sequence evolution [36].

Sequence data suffers from some limitations in phylogenetic reconstruction. The relatively fast pace of mutation in many regions of the genome results in *homoplasy* (multiple substitutions at the same position), leaving no trace in modern organisms of the actual series of events. Different regions of the genome (e.g. different genes), may not follow the same evolutionary path as the organism: this is known as the "gene tree species tree" problem. The problem of multiple sequence alignment between distant sequences is still poorly solved by computational approaches, and deep evolutionary histories are hard to reconstruct from sequence data.

1.1.2 Whole-genome data (at the level of syntenic blocks)

In whole-genome data, each chromosome of the genome is represented by an ordered list of identifiers, each identifier referring to a syntenic block or, more commonly, to a member of a gene family. (In the following, we shall use the word "gene" in a broader sense to denote elements of such orderings and refer to such orderings as "gene orders".) Variations in the placement of homologous genes, as well as variations in gene content and multiplicity, among organisms can then be analyzed. Such data is of great interest to evolutionary biologists, but also to comparative genomicists and to any researcher interested in understanding evolutionary changes in pathogens, crop plants, and, more generally, to anyone working in biomedical research. Evolutionary events that affect the gene order of genomes include various rearrangements, which affect only the order, and gene duplications and losses, which affect both the content and, indirectly, the order. Rearrangements themselves include inversion, transposition, block exchange, circularization and linearization, all of which act on a single chromosome, and translocation, fusion, and fission, which act on two chromosomes.

The use of whole-genome data is attractive in phylogenetic reconstruction. Genome rearrangements, gene duplications and losses are 'rare genomic events' and enable us to trace deep evolutionary history. The entire genome is studied at once as a single character, and the very large set of states for the genome is unlikely to give rise to homoplasy. The whole-genome data reflects organismal evolution, not the evolution of single genes, thereby avoiding the gene tree v.s. species tree problem.

1.2 Methods for phylogenetic reconstruction

Methods of phylogenetic reconstruction attempt to reverse a given model of evolution, given the data in modern organisms. There are three main types of methods, *distance-based*, *parsimony-based*, and *likelihood-based*.

1.2.1 Distance-based methods

These methods first estimate evolutionary distances between each pair of taxa, then use only the matrix of pairwise distances to reconstruct the phylogeny. The distances can be estimated as counts of the number of evolutionary events between two given taxa. Thus the estimation of pairwise distances must be done with respect to a chosen model of evolution. Since the true distance, that is, the actual number of changes that took place during the course of evolution, is not something we can compute, researchers have used a two-stage process, in which a well-defined measure is first computed (such as an edit distance, that is, the smallest number of evolutionary changes – from a defined set – needed to transform one into the other), then a statistical model of evolution is used to infer an estimate of the true distance by deriving the effect of a given number of changes in the model on the computed measure and (algebraically or numerically) inverting the derivation to produce a maximum-likelihood estimate of the true distance under the model. This second step is often called a distance correction and has long been used for sequence (DNA) data [107] as well as, more recently, for whole-genome data [76]. Once all pairwise distances have been computed, methods such as Neighbor-Joining [90] or FastME [17] can be used to reconstruct phylogeny.

1.2.2 Parsimony-based methods

These methods seek the tree and internal data that minimize the total number of evolutionary events needed to produce the leaves from a common ancestor. The total number of evolutionary events of a tree is the sum of its edge lengths where each edge length denotes the (edit) distance between the nodes at the two ends of the edge. In the case of sequence data, the general problem of finding the most parsimonious (MP) tree is provably NP-hard [15]. Current approaches are heuristics based on iterative improvement techniques, e.g., in MEGA [56], PAUP* [106], Phylip [31], and TNT [37]. With whole-genome data, the parsimony problem requires the inference of "ancestral" genomes at internal nodes of the candidate trees, which is NP-hard even for the median problem, a tree of three given genomes [12, 83]. Sankoff proposed to use the median problem in an iterative manner to refine ancestral genomes [92]; this approach was later improved in tools like GRAPPA [77, 110] and MGR [8].

1.2.3 Likelihood-based methods

Maximum likelihood (ML) methods assume a model of evolution, and aim to find the tree and associated model parameters, that maximize the probability of producing the given data. ML methods thus depend explicitly on the assumed model of evolution. ML is usually much more computationally expensive than MP, since ML has to estimate model parameters and search the best tree through tree space simultaneously [102]. Efficient heuristics exist for sequence data, e.g., PhyML [38] and RAxML [100]; a first attempt at using ML for whole-genome data appeared last year [44]. Bayesian methods assume a prior probability distribution of the possible trees, use a biased random walk through the tree space and estimate the posterior probability of trees given the data. The standard implementation is to use Markov Chain Monte Carlo (MCMC) approach, notably in tools MrBayes [45] for sequence data and a preliminary framework for whole-genome data [57].

In addition, meta-methods are used to scale up any of these methods in a divide-and-conquer way. They usually decompose the input dataset into overlapping subsets, reconstruct a tree for each subset, and combine those small trees to produce a complete tree for the original dataset. The most successful one is the *Disk-covering method* (DCM) [46], which improves both the speed and accuracy of existing approaches by carefully decomposing the dataset [48, 88].

1.3 Handling whole-genome data

In spite of many compelling reasons for using whole-genome data in phylogenetic reconstruction, practice to date has continued to use selected sequences of moderate length using nucleotide-, amino acid-, or codon-level models. Previous tools for reconstructing whole-genome phylogenies suffered from serious problems, usually combinations of oversimplified models, poor accuracy, poor scaling, lack of robustness against errors in the data, and lack of statistical assessment procedures.

Genomic rearrangements have been studied since the beginnings of modern genetics (starting in the 1920s with the classic work [103,104]) and models for such rearrangements have been the subject of many papers over the last 20 years [34], notably, the double-cut-and-join model (DCJ) [6, 119], which has formed the basis for much of the algorithmic research on whole-genome data over the last few years. However, none of the existing models predicts the evolution of genomic organization into circular unichromosomal genomes (as in most prokaryotes) and linear multichromosomal genomes (as in most eukaryotes). In addition, most of these models do not support gene duplications and losses alongside rearrangements; yet duplications and losses may be more common in evolutionary history than rearrangements, and moreover, they themselves cause apparent rearrangements.

The assessment of phylogenies built from whole-genome data has not been properly addressed to date. The standard method used in sequence-based phylogenetic inference is the bootstrap [23, 30], but it relies on a large number of homologous characters that can be resampled [30]; yet in the case of rearrangements, the entire genome is a single character. Alternatives such as the jackknife suffer from the same problem, while likelihood tests [3, 38] cannot be applied in the absence of well established probabilistic models.

Maximum-likelihood approaches have been successfully applied to phylogenetic inferences for aligned sequences, but such applications remain primitive for whole-genome data. It was not until last year that the first successful attempt to use ML reconstruction based on whole-genome data was published [44]; results from this study on bacterial genomes were promising, but somewhat difficult to explain, while the method is too time-consuming to handle eukaryotic genomes. A preliminary implementation of Bayesian methods has yielded some promising results, but was tested on just a few datasets [57].

1.4 Contributions in this dissertation

All the work presented in this dissertation has been accomplished by close collaboration with Bernard Moret. We have included only a part of our published research, the part where we played the lead role, spanning from models and distance estimation on whole-genome evolution to phylogenetic reconstruction with bootstrapping from whole-genome data. (the collaborations are mentioned in the following subsections)

1.4.1 Models and distance estimation on whole-genome evolution

(This is joint work with Vaibhav Rajan and Krister Swenson)

We present a method to estimate the true evolutionary distance between two genomes under the ‘double-cut-and-join’ (DCJ) model [6, 119] of genome rearrangements, a model under which a single multichromosomal operation accounts for all genomic rearrangement events: inversion, transposition, translocation, block interchange and chromosomal fusion and fission. Our method relies on a simple structural characterization of a genome pair and is both analytically and computationally tractable. We provide experimental results on a wide variety of genome structures to exemplify the very high accuracy (and low variance) of our estimator. The estimator also describes the asymptotic behavior of genome structure under the DCJ model, which motivates us to refine the DCJ model to account for biological constraints. The new evolutionary model introduces a single modification to the classic DCJ model, and integrates gene duplications and losses. Through these changes, it becomes the first mathematical model to preserve the structural dichotomy in genomic organization (1-2 circular chromosomes vs. several larger linear chromosomes) between most prokaryotes and most eukaryotes. These models and associated distance estimators on whole-genome evolution provide a basis for studying facets of possible mechanisms of evolution through simulation and application to real genomes.

1.4.2 Distance-based reconstruction with bootstrapping from whole-genome data

(This is joint work with Vaibhav Rajan)

We propose a new approach to the assessment of distance-based phylogenetic inference from whole-genome data; our approach combines features of the jackknife and the bootstrap and remains nonparametric. For each feature of our method, we give an equivalent feature in the sequence-based framework; we also present the results of extensive experimental testing, in both sequence-based and whole-genome-based frameworks. Through the feature-by-feature comparison and the experimental results, we show that our bootstrapping approach is on par with the classic phylogenetic bootstrap used in sequence-based reconstruction, and we establish the clear superiority of the classic bootstrap and of our corresponding new approach over proposed variants. We test our approach on a small dataset of mammalian genomes, verifying that the support values match current thinking about the respective branches. Our method is the first to provide a standard of assessment to match that of the classic phylogenetic bootstrap for aligned sequences, and thus makes it possible to conduct phylogenetic analyses on whole genomes with the same degree of confidence as for analyses on aligned sequences.

1.4.3 Maximum-likelihood reconstruction from whole-genome data

(This is joint work with Fei Hu and Jijun Tang)

We propose a maximum-likelihood approach to phylogenetic analysis from whole-genome data, in combination with our novel bootstrap scheme. Our approach uses a model that includes both rearrangements and duplications and losses; it is robust against common assembly errors; it supports bootstrapping and other standard statistical tests; it returns highly accurate trees in all our tests under a very wide variety of conditions; and it scales as well as approaches based on sequence data. The results of extensive testing on simulated data show that our approach returns very accurate results very quickly. In particular, we analyze of a 68-taxon collection of eukaryotic genomes [65], ranging from parasitic unicellular organisms with simple genomes to mammals and from around 3000 genes to over 40000 genes; the analysis, including bootstrapping, takes just 3 hours on a desktop system and returns a tree in agreement with all well supported branches, while also suggesting resolutions for some disputed placements.

Overall, we demonstrate that whole-genome data carries a very strong and robust phylogenetic signal and thus can form the basis for highly accurate phylogenetic analysis. While tools designed earlier were promising, with our new techniques described in this work, one can reconstruct accurate phylogenies from whole-genome data to an extent that was not possible before.

Chapter 2

Models and distance estimation on whole-genome evolution

The ordering and strandedness of genes on each chromosome of many organisms have become available, with many more added every year. Using this information, one can represent a genome as a collection of chromosomes, each of which is a linear or circular sequence of gene identifiers. Variations in the placement of the same genes, as well as variations in gene content and multiplicity, among organisms can then be analyzed. This data is of great interest to evolutionary biologists, but also to comparative genomicists and to any researcher interested in understanding evolutionary changes in pathogens. In the past ten years, there has been a large increase in work done on analyzing such data [74].

Perhaps the most basic requirement in the analysis of such data is the ability to estimate the amount of evolutionary change between two genomes—that is, to compute a pairwise *evolutionary distance*. Since the *true distance*, that is, the actual number of changes in the gene order and content that took place during the course of evolution, is not something we can compute, researchers have used a two-stage process, in which a well defined measure is first computed (such as an *edit distance*, that is, the smallest number of evolutionary changes—from a defined set—needed to transform one genome into the other), then a statistical model of evolution is used to infer an estimate of the true distance by deriving the effect of a given number of changes in the model on the computed measure and (algebraically or numerically) inverting the derivation to produce a maximum-likelihood estimate of the true distance under the model. This second step is often called a distance "correction" and has long been used for sequence (DNA) data [108] as well as, more recently, for gene-order data [73, 75, 115, 117].

The measures commonly used in the first step (edit distances, synteny measures, etc.) are bounded and typically reflect only the endstate of an evolutionary process, whereas the true evolutionary distance can be arbitrarily large. Thus these first-step measures typically underestimate the true distance, by an amount that grows quickly as the true distance grows large. This is an aspect of the problem of *saturation*, in which the evolutionary process may take a convoluted path to its endstate, possibly even undoing earlier changes along the way. For very small distances, the problem does not arise, while, for extremely large ones, the problem is essentially insurmountable, as the variance of any estimate will be huge. For most distance values, however, one can view the goal of distance correction as postponing the onset of saturation, that is, making it possible to deliver an accurate estimate of the true distance up to as large a value as possible.

Evolutionary events that affect the gene order of genomes include a number of rearrangements, which affect only the order, as well as gene duplication and loss, which affect both the gene content and, indirectly, the order. Handling both together has proved challenging [70, 105]. Rearrangements themselves include inversion, transposition, and block exchange, which act on a single chromosome, and translocation, fusion, and fission, which act on two chromosomes. Inversion, translocation, fusion, and fission were characterized by Hannenhalli and Pevzner [39, 40], while edit distances for these operations can be computed in linear time [4]. Sorting by transpositions has been proved to be NP-complete [9]. Efforts at unifying some of these operations in a statistical framework have had some success [18]. However, all these rearrangement operations are recently defined and studied a unifying operation in one or two steps: the so-called "double-cut-and-join", or DCJ, operation [119]. Bergeron *et al.* subsequently generalized the DCJ operation and showed how to compute an edit distance for it (assuming that every operation has unit cost) in linear time with a simple formula [6].

In Section 2.1, we address the problem of estimating a true evolutionary distance under the DCJ model of evolution, assuming no change in gene content and a uniform distribution of all possible DCJ events—the same simplifying assumptions used to date in all rearrangement analyses. In Section 2.2, we refine the DCJ model and propose a new evolutionary model which respects the dichotomy between prokaryotic and eukaryotic genomes and which takes gene duplications and losses into account. Using this new evolutionary model, we develop a statistically based method to estimate the true evolutionary distance in terms of the actual number of rearrangements, gene duplications, and gene losses.

2.1 Estimating true evolutionary distances under the DCJ model

Our estimate is in the style of the IEBP estimate for the true inversion distance for a single chromosome [115, 117], in that it does not require computing an edit distance, but only a simple count of shared gene adjacencies (or, equivalently, breakpoints, as in the work of Sankoff *et al.* [92, 93]) and chromosome endpoints. We characterize the asymptotic behavior of genome structure under the uniform DCJ model and present experimental results showing that our estimates are very precise, and exhibit very little variance, under both realistic and extreme parameter settings.

2.1.1 Preliminaries on whole-genome data and the DCJ model

A gene is a stranded sequence of DNA that starts with a tail and ends with a head. The tail of a gene a is denoted by a^t and its head by a^h. We write $+a$ ($a^t \rightarrow a^h$) if gene a is transcribed from $3'$ to $5'$ and write $-a$ ($a^h \rightarrow a^t$) otherwise. We are interested, not in the strand of one single gene, but in the connection of two consecutive genes in one chromosome. Due to different strandedness, two consecutive genes b and c can be connected by one *adjacency* of the following four types, $\{b^t, c^t\}$, $\{b^h, c^t\}$, $\{b^t, c^h\}$ and $\{b^h, c^h\}$. If gene d lies at one end of a linear chromosome, the we have a singleton set, $\{d^t\}$ or $\{d^h\}$, called *telomere*.

In the simplest case, we assume equal gene content and no duplicate gene. A *genome* is then represented as a set of adjacencies and telomeres such that the tail or the head of any gene appears in exactly one adjacency or telomere. For example, the genome G illustrated in Figure 2.1, composed of two linear chromosomes, $(+a, -c, -f)$ and $(+e)$, and one circular chromosome $(+b, +d)$, can be represented by the following set of adjacencies and telomeres: $\{\{a^t\}, \{a^h, c^h\}, \{c^t, f^h\}, \{f^t\}, \{b^h, d^t\}, \{d^h, b^t\}, \{e^t\}, \{e^h\}\}$.

The number of adjacencies and telomeres in one genome only captures the number of linear chromosomes: k adjacencies from circular chromosomes could come from a single circular chromosome

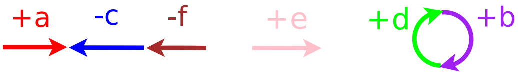


Figure 2.1: A very small genome G

of size k or from k circular chromosomes of one gene each, or any other combination. In particular, every genome on n genes made entirely of circular chromosomes has the same number of adjacencies and telomeres.

The double-cut-and-join operation, in the formulation of [6], can model all classical rearrangements: inversion, translocation, fusion, fission, transposition and block interchange. In that formulation, a DCJ operation makes a pair of cuts in the chromosomes and reglues the cut ends on two adjacencies or telomeres (which can be in the same chromosome or in different chromosomes), giving rise to four cases:

1. A pair of adjacencies $\{i^u, j^v\}$ and $\{p^x, q^y\}$ can be replaced by the pair $\{i^u, p^x\}$ and $\{j^v, q^y\}$ or by the pair $\{i^u, q^y\}$ and $\{j^v, p^x\}$.
2. An adjacency $\{i^u, j^v\}$ and a telomere $\{p^x\}$ can be replaced by the adjacency $\{i^u, p^x\}$ and telomere $\{j^v\}$ or by the adjacency $\{j^v, p^x\}$ and telomere $\{i^u\}$.
3. A pair of telomeres $\{i^u\}$ and $\{j^v\}$ can be replaced by the adjacency $\{i^u, j^v\}$.
4. An adjacency $\{i^u, j^v\}$ can be replaced by the pair of telomeres $\{i^u\}$ and $\{j^v\}$.

Theorem 2.1.1. *Let G be a genome with n genes, n_1 adjacencies, and n_2 telomeres. If m is the number of the different possible DCJ operations on G, we can write*

$$\begin{aligned} n &= n_1 + \frac{n_2}{2} \\ m &= n_1^2 + 2n_1n_2 + \frac{1}{2}n_2^2 - \frac{1}{2}n_2 \\ n^2 &\leq m \leq 2n^2 - n \end{aligned}$$

Proof. G has n genes and thus $2n$ tails and heads of genes; as the tail or the head of any gene appears in exactly one adjacency or telomere, we have

$$2n = 2n_1 + n_2 \tag{2.1}$$

Now consider the four cases of DCJ operations:

1. There are $\binom{n_1}{2}$ ways to select two adjacencies and 2 possible DCJ operations for each such choice, for a total of $\binom{n_1}{2} \times 2$ operations.
2. There are $n_1 \times n_2$ ways to select one adjacency and one telomere and 2 possible DCJ operations for each combination, for a total of $n_1 \times n_2 \times 2$ operations.
3. There are $\binom{n_2}{2}$ ways to select two telomeres and 1 possible DCJ operation for each such choice, for a total of $\binom{n_2}{2}$ operations.
4. There are n_1 different ways to select one adjacency and 1 possible DCJ operation for each such choice, for a total of n_1 operations.

Thus the total number of possible DCJ operations is

$$m = n_1^2 + 2n_1n_2 + \frac{1}{2}n_2^2 - \frac{1}{2}n_2$$

Combining this result with formula (2.1), we get

$$m = -\frac{1}{4}n_2^2 + (n - \frac{1}{2})n_2 + n^2$$

Now we also have $0 \leq n_2 \leq 2n$, and so we can write

$$n^2 \leq m \leq 2n^2 - n$$

□

2.1.2 True distance estimation under the DCJ model

An overview of our technique for estimating the true evolutionary distance

The problem of estimating the true evolutionary distance under DCJ model is defined as follows:
Input: The original genome G and the final genome G^F, two genomes on the same n genes represented as adjacencies and telomeres.
Output: An estimate of the actual number of DCJ operations that took place in the evolutionary history to transform G into G^F.

Based on the original genome G, for any genome G^* (of same gene content as G), we can divide the adjacencies and telomeres of G^* into four sets $SA(G^*)$, $ST(G^*)$, $DA(G^*)$ and $DT(G^*)$, where $SA(G^*)$ is the set of adjacencies of G^* that also appear in G, $ST(G^*)$ is the set of telomeres of G^* that also appear in G, $DA(G^*)$ is the set of adjacencies of G^* that do not appear in G, and $DT(G^*)$ is the set of telomeres of G^* that do not appear in G. Then we can calculate a vector $V_G(G^*) = (SA^*, ST^*, DA^*, DT^*)$ to represent the genome G^* based on G, where SA^*, ST^*, DA^* and DT^* are the cardinalities of the sets $SA(G^*)$, $ST(G^*)$, $DA(G^*)$ and $DT(G^*)$, respectively. (V_G may be viewed as producing a fingerprint of G^*.) Obviously, we have

$$2n = 2SA^* + ST^* + 2DA^* + DT^*$$

Let G^k be the genome obtained from $G = G^0$ by applying k randomly selected DCJ operations—under our model, the $(i+1)$st DCJ operation is selected from a uniform distribution of all possible DCJ operations on the current genome G^i. We can compute the vector $V_G(G^k) = (SA^k, ST^k, DA^k, DT^k)$ to represent the genome G^k with respect to G.

Now we will show that, given G, we can also produce the estimate $\widetilde{E}(V_G(G^k)) = (\widetilde{SA^k}, \widetilde{ST^k}, \widetilde{DA^k}, \widetilde{DT^k})$ for the expected vector $E(V_G(G^k))$ for any integer $k > 0$. We use $\widetilde{SA^k}$ to approximate the expected number of adjacencies present in both G and G^k. We compute SA^F from G and G^F. Our approach for estimating the true evolutionary distance is then to return the integer k that minimizes the difference $|SA^F - \widetilde{SA^k}|$.

Estimation of the expected vector after some number of random DCJ operations

We show how to estimate the expected vector $E(V_G(G^k))$ under our DCJ model for any integer $k > 0$.

Let G and G^k be as defined above; the vector for $G^0 = G$ is clearly just $V_G(G^0) = (n_1, n_2, 0, 0)$. We first show how to compute $E(V_G(G^1))$.

Theorem 2.1.2. *Let m be the number of possible DCJ operations applicable to G. We have $E(V_G(G^1)) = (SA^1, ST^1, DA^1, DT^1)$, where*

$$\begin{aligned}
SA^1 &= n_1 - \frac{2n_1^2 + 2n_1n_2 - n_1}{m} \\
ST^1 &= n_2 - \frac{2n_1n_2 + n_2^2 - n_2}{m} \\
DA^1 &= \frac{2n_1^2 - 2n_1 + 2n_1n_2 + \frac{1}{2}n_2^2 - \frac{1}{2}n_2}{m} \\
DT^1 &= \frac{2n_1n_2 + 2n_1}{m}
\end{aligned}$$

Proof. Write $V_G(G^0) = (SA^0, ST^0, 0, 0)$ and consider the four cases for DCJ operations.

1. When we select two adjacencies out of $SA(G^0)$, the number of possible DCJ operations is $\binom{SA^0}{2} \times 2$. Neither of the resulting adjacencies will be in G, so that every such operation reduces SA^0 by 2 and increase DA^0 by 2.

2. When we select one adjacency out of $SA(G^0)$ and one telomere out of $ST(G^0)$, the number of possible DCJ operations is $SA^0 \times ST^0 \times 2$. Neither of the resulting adjacency nor telomere will be in G, so that every such operation reduces both SA^0 and ST^0 by 1 and increases both DA^0 and DT^0 by 1.
3. When we select two telomeres out of $ST(G^0)$, the number of possible DCJ operations is $\binom{ST^0}{2}$. The resulting adjacency will not be in G, so that every such operation will reduce ST^0 by 2 and increase DA^0 by 1.
4. When we select one adjacency out of $SA(G^0)$, the number of possible DCJ operations is SA^0. Neither of the resulting telomeres will be in G, so that every such operation reduces SA^0 by 1 and increases DT^0 by 2.

Adding up the 4 cases and normalizing by the total m, we get

$$\begin{aligned}
SA^1 &= SA^0 + \frac{2\binom{SA^0}{2}}{m}\cdot(-2) + \frac{2SA^0ST^0}{m}\cdot(-1) + \frac{SA^0}{m}\cdot(-1) \\
&= SA^0 - \frac{2SA^{0^2} + 2SA^0ST^0 - SA^0}{m} \\
ST^1 &= ST^0 + \frac{SA^0\cdot ST^0\cdot 2}{m}\cdot(-1) + \frac{\binom{ST^0}{2}}{m}\cdot(-2) \\
&= ST^0 - \frac{2SA^0ST^0 + ST^{0^2} - ST^0}{m} \\
DA^1 &= 0 + \frac{\binom{SA^0}{2}\cdot 2}{m}\cdot 2 + \frac{SA^0\cdot ST^0\cdot 2}{m}\cdot 1 + \frac{\binom{ST^0}{2}}{m}\cdot 1 \\
&= \frac{2SA^{0^2} - 2SA^0 + 2SA^0ST^0 + \frac{1}{2}ST^{0^2} - \frac{1}{2}ST^0}{m} \\
DT^1 &= 0 + \frac{SA^0\cdot ST^0\cdot 2}{m}\cdot 1 + \frac{SA^0}{m}\cdot 2 \\
&= \frac{2SA^0ST^0 + 2SA^0}{m}
\end{aligned}$$

□

Let G^k be a genome obtained from G by applying k randomly selected DCJ operations and let G^{k+1} be the genome obtained from the genome G^k by applying one more randomly selected DCJ operation. We show how to calculate the expected value of $V_G(G^{k+1})$ given G^k and G.

Theorem 2.1.3. *Let $V_G(G^k) = (SA^k, ST^k, DA^k, DT^k)$ and let m_k be the number of possible DCJ operations on G^k. For conciseness, write $A^k = SA^k + DA^k$ (the number of adjacencies in G^k) and $T^k = ST^k + DT^k$ (the number of telomeres in G^k). Then we can write*

$$\begin{aligned}
m_k &= (A^k)^2 + 2(A^k)(T^k) + \frac{1}{2}(T^k)^2 - \frac{1}{2}(T^k) \\
E(V_G(G^{k+1})) &= (SA^{k+1}, ST_T^{k+1}, DA^{k+1}, DT^{k+1})
\end{aligned}$$

where we have

$$\begin{aligned}
SA^{k+1} &= SA^k + \frac{1}{m_k}[n_1 - 2SA^k(A^k + T^k)] \\
ST^{k+1} &= ST^k + \frac{1}{m_k}[n_2(T^k+1) - 2ST^k(A^k + T^k)] \\
DA^{k+1} &= DA^k + \frac{1}{m_k}[2SA^k(A^k + T^k) + \binom{T^k}{2} - (A^k) - n_1] \\
DT^{k+1} &= DT^k + \frac{1}{m_k}[2ST^k(A^k + T^k) - n_2(T^k+1) - 2\binom{T^k}{2} + 2(A^k)]
\end{aligned}$$

Proof. From Theorem 2.1.1, we have

$$m_k = (A^k)^2 + 2(A^k)(T^k) + \frac{1}{2}(T^k)^2 - \frac{1}{2}(T^k)$$

There are $n_1 - SA^k$ adjacencies in G that do not appear in G^k and they must fall into one the following 3 cases:

1. n_{AA} pairs with members in two different adjacencies in $DA(G^k)$.
2. n_{TT} pairs with members in two telomeres of $DT(G^k)$.
3. n_{AT} pairs with one member in $DA(G^k)$ and the other in $DT(G^k)$.

There also are $n_2 - ST^k$ telomeres in G that do not appear in G^k and so must be members of $DA(G^k)$.

Now we complete the proof by running through the possible cases. From the proof for Theorem 2.1.2, we have already covered 4 cases where adjacencies and telomeres were selected only from $SA(G^k)$ and $ST(G^k)$. The remaining 8 cases cover selections from $DA(G^k)$ and $DT(G^k)$ as well. In the last 5 of these 8 cases, the outcome of a particular operation in terms of adjacency and telomere counts is not fixed, but the total count over all possible operations can still be computed; we use the expression “recover” (an adjacency or a telomere) to indicate a case in which the count increases.

1. When we select one adjacency out of $SA(G^k)$ and another out of $DA(G^k)$, the number of possible DCJ operations is $SA^k \times DA^k \times 2$. Neither resulting adjacency will be in G, so that every such operation reduces SA^k by 1 and increases DA^k by 1.
2. When we select one adjacency out of $SA(G^k)$ and one telomere out of $DT(G^k)$, the number of possible DCJ operations is $SA^k \times DT^k \times 2$. Neither the resulting adjacency nor telomere will be in G, so that every such operation reduces SA^k by 1 and increases DA^k by 1.
3. When we select one telomere out of $ST(G^k)$ and one telomere out of $DT(G^k)$, the number of possible DCJ operations is $ST^k \times DT^k$. Neither the resulting adjacency nor telomere will be in G, so that every such operation reduces ST^k and DT^k by 1 and increases DA^k by 1.
4. When we select one telomere out of $ST(G^k)$ and one adjacency out of $DA(G^k)$, the number of possible DCJ operations is $ST^k \times DA^k \times 2$. The resulting adjacency will not be in G, while the resulting telomere can be in G or not. There are $ST^k \times (n_2 - ST^k)$ ways to recover one telomere out of $n_2 - ST^k$ telomeres in G that do not appear in G^k.
5. When we select two adjacencies out of $DA(G^k)$, the number of possible DCJ operations is $\binom{DA^k}{2} \times 2$. The two resulting adjacencies can be in G or not. There are n_{AA} ways to recover one adjacency out of $n_1 - SA^k$ adjacencies in G that do not appear in G^k.
6. When we select one one adjacency out of $DA(G^k)$ and one telomere out of D_T^k, the number of possible DCJ operations is $DA^k \times DT^k \times 2$. The resulting adjacency and telomere can be in G or not. There are $DT^k \times (n_2 - ST^k)$ ways to recover one telomere out of $n_2 - ST^k$ telomeres in G that do not appear in G^k and n_{AT} ways to recover one adjacency out of $n_1 - SA^k$ adjacencies in G that do not appear in G^k.

7. When we select one adjacency out of $DA(G^k)$, the number of possible DCJ operations is DA^k. The two resulting telomeres can be in G or not and there are $n_2 - ST^k$ ways to recover one telomere out of $n_2 - ST^k$ telomeres in G that do not appear in G^k.
8. When we select two telomeres out of $DT(G^k)$, the number of possible DCJ operations is $\binom{DT^k}{2}$. The resulting adjacency can be in G or not and there are n_{TT} ways to recover one adjacency out of $n_1 - SA^k$ adjacencies in G that do not appear in G^k.

Adding up the 12 cases and normalizing by the total m_k, we get

$$\begin{aligned}
SA^{k+1} &= SA^k + \frac{1}{m_k}[n_1 - 2SA^k(A^k + T^k)] \\
ST^{k+1} &= ST^k + \frac{1}{m_k}[n_2(T^k+1) - 2ST^k(A^k + T^k)] \\
DA^{k+1} &= DA^k + \frac{1}{m_k}[2SA^k(A^k + T^k) + \binom{T^k}{2} - (A^k) - n_1] \\
DT^{k+1} &= DT^k + \frac{1}{m_k}[2ST^k(A^k + T^k) - n_2(T^k+1) - 2\binom{T^k}{2} + 2(A^k)]
\end{aligned}$$

□

Given G^0, we estimate $E(V_G(G^k))$ for $k > 0$ by iterating k times the matching formula in Theorem 2.1.3, and every time we identify $E(V_G(G^{k-1}))$ with the actual vector $V_G(G^{k-1})$.

Corollary 2.1.4. *Let G be one genome on n genes, the estimated vector $\widetilde{E}(V_G(G^i)) = (\widetilde{SA^i}, \widetilde{ST^i}, \widetilde{DA^i}, \widetilde{DT^i})$ for all integers i $(0 \le i \le k)$ can be computed in $O(k)$ time.*

2.1.3 Experimental results

We now present experimental results on the accuracy of our estimation of the expected vector after a given number of random DCJ operations and on the quality of our estimator for the true evolutionary distance (in terms of the actual number of DCJ operations). Our experiments all start with an original genome, G, with some chosen number of linear and circular chromosomes of various sizes; this genome is subjected to a prescribed number k of DCJ operations chosen uniformly at random to obtain a final genome G^k. We vary k from one to six times the number of genes—very large values in evolutionary terms. For each choice of parameters, we generate $10,000$ runs to obtain a tight estimate of variance. We compute the vector representations for all intermediate genomes and then use our method to estimate the evolutionary distance. We run tests on a large variety of initial genomes and present here results on three initial genomes, all meant to resemble real organismal genomes: (a) $25,000$ genes and 25 linear chromosomes; (b) $10,000$ genes and 5 linear chromosomes; and (c) $1,000$ genes and 1 circular chromosome—the first two examples match metazoan genomes, the last matches a small bacterial genome.

Accuracy of the expected vector after k random DCJ operations

We study the behavior of our estimation $\widetilde{E}(V_G(G^k))$ by comparing its prediction to the sample mean for $E(V_G(G^k))$, as computed from our $10,000$ trials. We compute the mean absolute difference for SA, ST, DA, and DT between our estimation $\widetilde{E}(V_G(G^k))$ and each experimental vector $V_G(G^k)$ in every single run for genomes (a), (b), and (c) and show the results in Figure 2.2.

The sum of absolute difference of entries in the vector on the larger genomes never exceeds 0.5% (as a percentage of the sum of entries in the vector) and is typically well below 0.25%; even on the smaller genome, the difference does not exceed 2% and is typically below 1%.

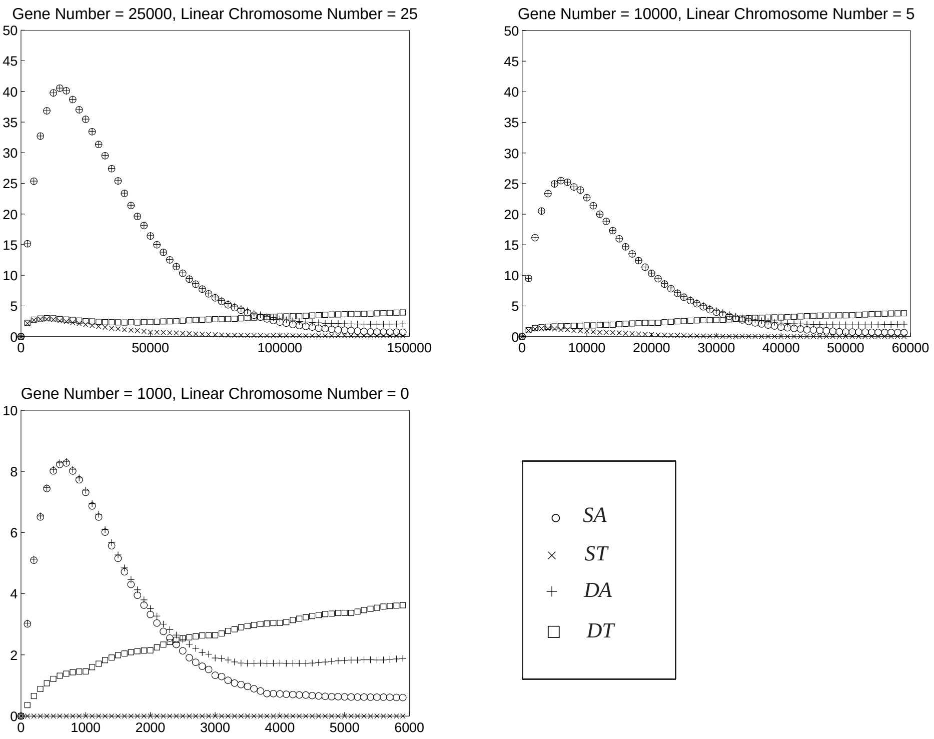


Figure 2.2: The mean absolute difference for SA, ST, DA and DT between our estimation $\widetilde{E}(V_G(G^k))$ and each experimental vector $V_G(G^k)$ as a function of the actual number of DCJ operations.

Accuracy of the estimation of the actual number of DCJ operations

We want to study the threshold of saturation of our estimator in addition to its accuracy; in order to do that, we create simulations with controlled numbers of DCJ operations and set up a threshold for correction in the estimation procedure. Specifically, we choose a number between 1 and some upper bound B as the actual number of DCJ operations; B is chosen to be the smallest integer k that makes the expected value $\widetilde{SA^k}$ smaller than 2, a point at which there are almost no shared adjacencies left. For genomes (a), (b) and (c), the corresponding upper bounds are $121,621$, $44,047$, and $3,253$, respectively. We use the smallest integer r that causes the expected value $\widetilde{s_A^r}$ to become smaller than $\frac{1}{2}$ as an upper limit on the maximum number of DCJ operations in the evolutionary history. Finally, if we have $s_A^F = 0$, we set k (the value normally chosen to minimize $|SA^F - \widetilde{SA^k}|$) to this upper limit r. For genomes (a), (b) and (c), r has values $211,442$, $81,329$, and $6,398$, respectively.

Figure 2.3 shows the mean and standard deviation for the actual number of DCJ operations estimated by the edit DCJ distance and by our approach. These figures indicate that, as expected, the edit DCJ distance underestimates, often severely, the actual number of events. In contrast, our approach provides highly accurate estimates, with very small variance.

We also study the mean absolute difference between the actual number of DCJ operations and our estimator for genomes (a), (b) and (c). The results are shown in Table 2.1.

The estimates are highly accurate (even for small genomes) up to surprisingly large numbers of events. Rearrangement events fall under the category of “rare genomic events” [87], yet our estimator works well even for what would be considered common events.

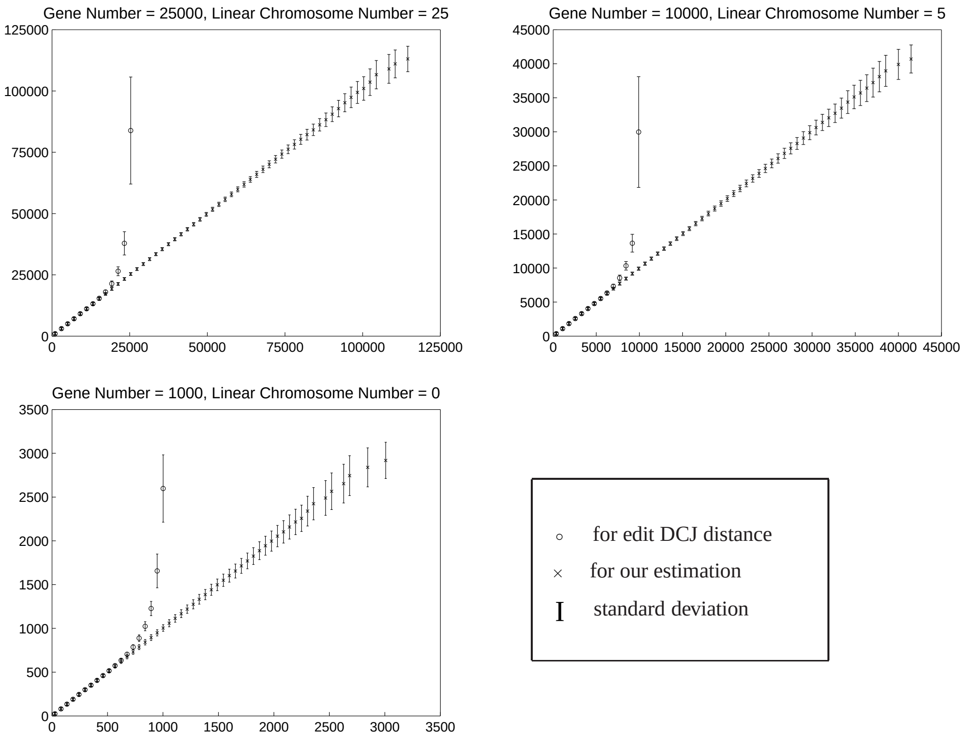


Figure 2.3: Mean and standard deviation plots for the actual number of DCJ operations (y axis) vs. the edit DCJ distance and our estimator (x axis). The datasets are divided into 60 bins according to their x-coordinate values.

Table 2.1: *The mean absolute difference between actual number of DCJ operations and our estimation.*

	actual number of DCJ operations		
# genes	# genes ×1	# genes ×2	# genes ×3
25,000	131.0 (0.5%)	447.5 (0.9%)	1280.2 (1.7%)
10,000	83.9 (0.8%)	282.0 (1.4%)	819.4 (2.7%)
1,000	27.2 (2.7%)	93.6 (4.7%)	441.8 (14.7%)

2.1.4 Discussion

From Figure 2.3, our approach postpones the threshold of saturation (viewed as a number of DCJ operations) from well under the number of genes to at least three times the number of genes for all three example genomes. This large gain in accuracy should translate into much better phylogenetic reconstructions as well as more accurate genomic alignments.

There are two main assumptions made in this work: no gene duplication or loss; and uniform distribution of DCJ operations. Both are clearly unrealistic, so our ability to gauge their effect on model predictions is crucial to future model refinements.

For instance, the DCJ model requires that a chromosomal fission that creates a new small circular chromosome be immediately followed by a chromosomal fusion that re-absorbs this small circular chromosome, thereby causing a block exchange within the original chromosome and treating the extra circular chromosome as a transient artifact [119]. Since circular chromosomes do not arise in organisms with a number of linear chromosomes, a similar constraint would strongly reduce the incidence of fission. A similar type of constraint could be used for prokaryotic genomes, which normally consist of a single circular chromosome. Evidence that paracentric rearrangements are more common than pericentric ones, at least in two *Drosophila* species [121], and that short inversions are more common than long ones, in some prokaryotes and in the aforementioned *Drosophila* [58, 121], can also be reflected into additional constraints on the DCJ model. Any additional constraint naturally creates complications, but we expect that at least a few natural constraints can be handled within the framework described here.

Since the DCJ operation regroups all rearrangements studied to date, and since our results point to one way in which the behavior of this model can be studied for various constraints (such as where the cuts can be made), our results may shed light on the vexing issue of what constitutes a significant syntenic block in comparative genomics—an issue that has seen a lot of discussion over the last few years [13, 96].

2.2 Models and distance estimations under rearrangements, duplications, and losses

In the previous section, we present a statistically based method to estimate the true evolutionary distance between two genomes under the DCJ model. The DCJ model, however, is unrealistic in two major respects. First, if the two cuts are in the same chromosome, one of the two nontrivial rejoinings causes a fission, creating a new circular chromosome. However, circular chromosomes do not normally arise in organisms with linear chromosomes, and prokaryotic genomes normally consist of a single circular chromosome. Nor can this form of rejoining be forbidden as, without it, DCJ simply reduces to inversion. Secondly, DCJ is a model of rearrangements: it does not take into account evolutionary events that alter the gene content, such as duplications and losses.

Of these two problems, the first has not been seriously addressed: the model we present here is, to the best of our knowledge, the first model that naturally preserves the dichotomy between prokaryotic and eukaryotic genomes. While gene (or segment) duplications and losses have long been studied by geneticists and molecular biologists, little work has been done to date on integrating them with rearrangements in a unified model. El-Mabrouk [24] gave an exact algorithm to compute edit distances for inversions and losses and also a heuristic to approximate edit distances for inversions, losses, and nonduplicating insertions (all of her results assume that genes cannot be duplicated). More recently, Yancopoulos and Friedberg [120] gave an algorithm to compute edit distances under deletions, insertions, duplications, and DCJ operations, under the constraint that each deletion can only remove a single gene. These and other approaches targeted the edit distance, not the true evolutionary distance. Swenson *et al.* [105] gave an algorithm to approximate the true evolutionary distance under

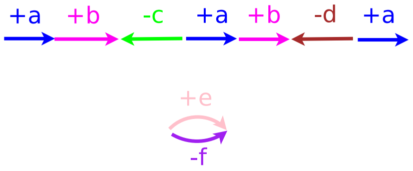


Figure 2.4: A very small genome G

deletions, insertions, duplications, and inversions for unichromosomal genomes and showed good results under simulations and for small-scale phylogenetic reconstruction. Rearrangements, duplications and losses have also been addressed in the framework of ancestral reconstruction [67, 82]. All of these approaches have focussed on parsimony criteria and have used pre-assigned weights for the various operations.

We propose a new evolutionary model which respects the dichotomy between prokaryotic and eukaryotic genomes and which takes gene duplications and losses into account, and develop a statistically based method to estimate the true evolutionary distance under our new model.

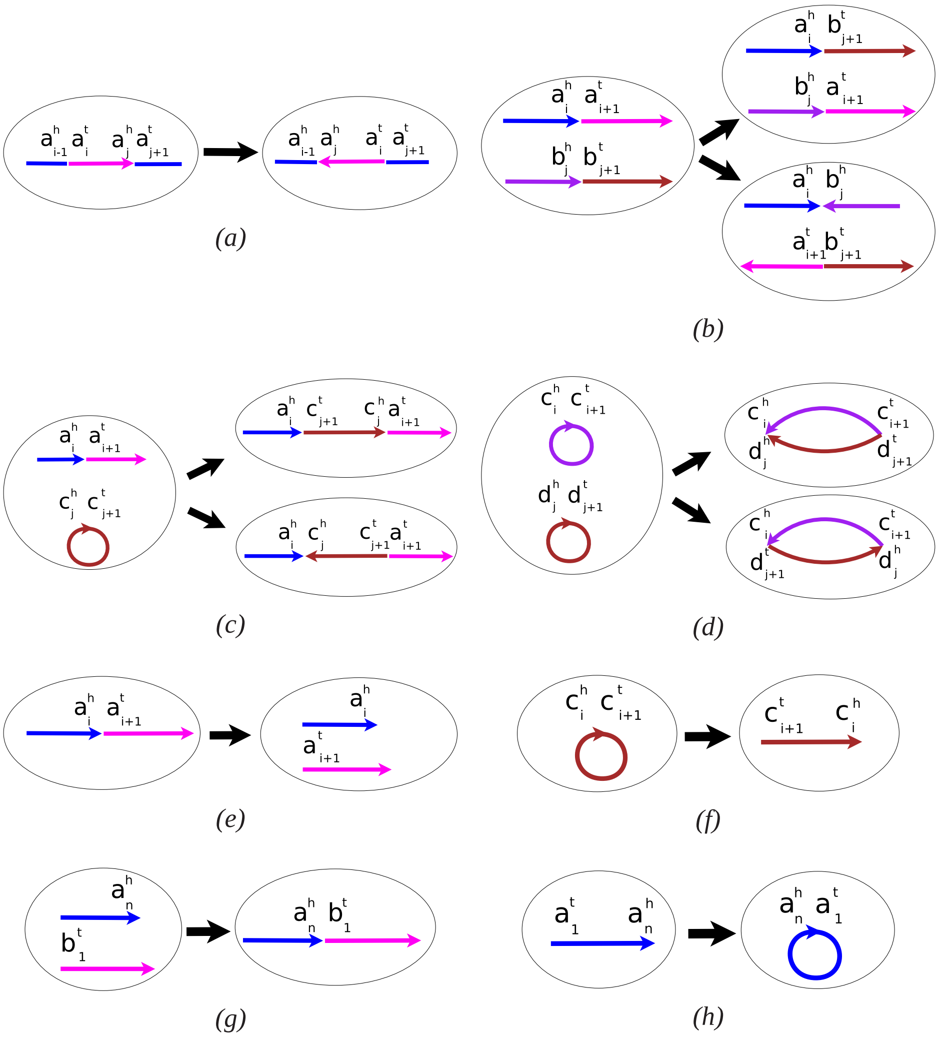
2.2.1 Preliminaries on gene-order data and the new evolutionary model

We denote the tail of a gene g by g^t and its head by g^h. We write $+g$ to indicate an orientation from tail to head ($g^t \to g^h$), $-g$ otherwise ($g^h \to g^t$). Two consecutive genes a and b can be connected by one *adjacency* of one of the following four types: $\{a^t,b^t\}$, $\{a^h,b^t\}$, $\{a^t,b^h\}$, and $\{a^h,b^h\}$. If gene c lies at one end of a linear chromosome, then we also have a singleton set, $\{c^t\}$ or $\{c^h\}$, called a *telomere*. A *genome* can then be represented as a multiset of genes together with a multiset of adjacencies and telomeres. For example, the toy genome in Figure 2.4, composed of one linear chromosome, $(+a,+b,-c,+a,+b,-d,+a)$, and one circular one, $(+e,-f)$, can be represented by the multiset of genes $\{a,a,a,b,b,c,d,e,f\}$ and the multiset of adjacencies and telomeres $\{\{a^t\}, \{a^h,b^t\}, \{b^h,c^h\}, \{c^t,a^h\}, \{a^h,b^t\}, \{b^h,d^h\}, \{d^t,a^h\}, \{a^h\}, \{e^h,f^h\}, \{e^t,f^t\}\}$. Because of the duplicated genes, there is no one-to-one correspondence between genomes and multisets of genes, adjacencies, and telomeres. For example, the genome composed of one linear chromosome, $(+a,+b,-d,+a,+b,-c,+a)$ and one circular one $(+e,-f)$ would have the same multisets of genes, adjacencies and telomeres as that in Figure 2.4.

In the new evolutionary model, a genomic change is one of a gene duplication, a gene loss, or a genome rearrangement, so that there are two parameters: the probability of occurrence of a gene duplication, p_d, and the probability of occurrence of a gene loss, p_l—the probability of occurrence of a rearrangement is then just $p_r = 1 - p_d - p_l$. The next event is chosen from the three categories according to these parameters.

For rearrangements, we select two elements uniformly *with replacement* from the multiset of all adjacencies and telomeres and then decide which rearrangement event we apply to these two elements. Compared to the DCJ model, the new model assigns a specific probability to each operation and forbid the one operation that creates circular intermediates. Thus we have eight cases in all (refer to Figure 2.5). For each case, we apply the intuitive interpretation in terms of replacing sets of adjacencies and telomeres suggested by Bergeron *et al.* [5, 6].

Select two different adjacencies, or one adjacency and one telomere, in the same chromosome (Figure 2.5a). For example, select two different adjacencies $\{a^h_{i-1},a^t_i\}$ and $\{a^h_j,a^t_{j+1}\}$ on one linear chromosome $A = (a_1 \ldots a_{i-1}a_i \ldots a_j a_{j+1} \ldots a_n)$. Reversing all genes between a_i and a_j yields $(a_1 \ldots a_{i-1}\text{-}a_j \ldots \text{-}a_i a_{j+1} \ldots a_n)$. Two adjacencies, $\{a^h_{i-1},a^t_i\}$ and $\{a^h_j,a^t_{j+1}\}$, are replaced by two others, $\{a^h_{i-1},a^h_j\}$ and $\{a^t_i,a^t_{j+1}\}$. This operation causes an inversion. (It is in this case


a h i-1 a t i a h j a t j+1
a h i-1 a h j a t i a t j+1
a h i a t i+1
b h j b t j+1
a h i b t j+1
b h j a t i+1
a h i b h j
a t i+1 b t j+1
a h i a t i+1
c h j c t j+1
a h i c t j+1 c h j a t i+1
a h i c h j c t j+1 a t i+1
c h i c t i+1
d h j d t j+1
c h i c t i+1
d h j d t j+1
c h i c t i+1
d t j+1 d h j
a h i a t i+1
a h i
a t i+1
c h i c t i+1
c t i+1 c h i
a h n
b t 1
a h n b t 1
a t 1 a h n
a h n a t 1

(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

Figure 2.5: Possible rearrangements.

that we forbid the creation of a new circular chromosome through fission, which would use the same choices of adjacencies, but rejoin the pieces differently.)

Select two adjacencies, or one adjacency and one telomere, in two linear chromosomes (Figure 2.5b). For example, select two adjacencies, $\{a_i^h, a_{i+1}^t\}$ from one linear chromosome $A = (a_1 \dots a_i a_{i+1} \dots a_n)$ and $\{b_j^h, b_{j+1}^t\}$ from another linear chromosome $B = (b_1 \dots b_j b_{j+1} \dots b_m)$. Now exchange the two segments between these two chromosomes C and D. There are two possible outcomes, $(a_1 \dots a_i b_{j+1} \dots b_m)$ and $(b_1 \dots b_j a_{i+1} \dots a_n)$ or $(a_1 \dots a_i\text{-}b_j \dots \text{-}b_1)$ and $(\text{-}b_n \dots \text{-}b_{j+1} a_{i+1} \dots a_n)$. Two adjacencies, $\{a_i^h, a_{i+1}^t\}$ and $\{b_j^h, b_{j+1}^t\}$, are replaced by $\{a_i^h, b_{j+1}^h\}$ and $\{a_{i+1}^t, b_j^t\}$ or $\{a_i^h, b_j^h\}$ and $\{a_{i+1}^t, b_{j+1}^t\}$. This operation causes a translocation.

Select two different adjacencies, or one adjacency and one telomere, in one circular chromosome and one linear chromosome (Figure 2.5c). For example, select two adjacencies, $\{a_i^h, a_{i+1}^t\}$ from one linear chromosome $A = (a_1 \dots a_i a_{i+1} \dots a_n)$ and $\{c_j^h, c_{j+1}^t\}$ one circular chromosome $C = (c_1 \dots c_j c_{j+1} \dots c_m)$. Now merge the circular chromosome C into the linear chromosome A. There are two possible outcomes, linear chromosomes $(a_1 \dots a_i c_{j+1} \dots c_m c_1 \dots c_j a_{i+1} \dots a_n)$ or
$(a_1 \dots a_i\text{-}c_j \dots \text{-}c_1\text{-}c_m \dots \text{-}c_{j+1} a_{i+1} \dots a_n)$. Two adjacencies, $\{a_i^h, a_{i+1}^t\}$ and $\{c_j^h, c_{j+1}^t\}$, are re-

placed by $\{a_i^h, c_{j+1}^h\}$ and $\{a_{i+1}^t, c_j^t\}$ or $\{a_i^h, c_j^h\}$ and $\{a_{i+1}^t, c_{j+1}^t\}$. This operation causes a fusion of a circular chromosome with a linear chromosome.

Select two adjacencies in two circular chromosomes (Figure 2.5d). For example, select two adjacencies, $\{c_i^h, c_{i+1}^t\}$ from one circular chromosome $C = (c_1 \dots c_i c_{i+1} \dots c_m)$ and $\{d_j^h, d_{j+1}^t\}$ from another circular chromosome $D = (d_1 \dots d_j d_{j+1} \dots d_n)$. Now merge these two circular chromosomes C and D into one new circular chromosome. There are two possible outcomes, circular chromosomes
$(c_1 \dots c_i d_{j+1} \dots d_m d_1 \dots d_j c_{i+1} \dots c_m)$ or $(c_1 \dots c_i\text{-}d_j \dots \text{-}d_1\text{-}d_m \dots \text{-}d_{j+1} c_{i+1} \dots c_m)$. Two adjacencies, $\{c_i^h, c_{i+1}^t\}$ and $\{d_j^h, d_{j+1}^t\}$, are replaced by $\{c_i^h, d_{j+1}^h\}$ and $\{c_{i+1}^t, d_j^t\}$ or $\{c_i^h, d_j^h\}$ and $\{c_{i+1}^t, d_{j+1}^t\}$. This operation causes a fusion of two circular chromosomes.

Select the same adjacency twice in one linear chromosome (Figure 2.5e). For example, select the adjacency $\{a_i^h, a_{i+1}^t\}$ twice from linear chromosome $A = (a_1 \dots a_i a_{i+1} \dots a_n)$. Then split C into two new linear chromosomes, $(a_1 \dots a_i)$ and $(a_{i+1} \dots a_n)$. The adjacency $\{a_i^h, a_{i+1}^t\}$ is replaced by two telomeres $\{a_i^h\}$ and $\{a_{i+1}^t\}$. This operation causes a fission of a linear chromosome.

Select the same adjacency twice in one circular chromosome (Figure 2.5f). For example, select the adjacency $\{c_i^h, c_{i+1}^t\}$ twice from circular chromosome $C = (c_1 \dots c_i c_{i+1} \dots c_m)$. Then linearize C into a linear chromosome, $(c_{i+1} \dots c_m c_1 \dots c_i)$. The adjacency $\{c_i^h, c_{i+1}^t\}$ is replaced by two telomeres $\{c_i^h\}$ and $\{c_{i+1}^t\}$. This operation causes a linearization of a circular chromosome.

Select two telomeres in two linear chromosomes (Figure 2.5g). For example, select telomeres $\{a_n^h\}$ and $\{b_1^t\}$ from two different linear chromosomes $A = (a_1 \dots a_i a_{i+1} \dots a_n)$ and $B = (b_1 \dots b_j\ b_{j+1} \dots b_m)$. Then concatenate these two linear chromosomes into a single new chromosome $(a_1 \dots a_i\ a_{i+1} \dots a_n b_1 \dots b_j b_{j+1} \dots b_m)$. Two telomeres, $\{a_n^h\}$ and $\{b_1^t\}$, are replaced by one adjacency $\{a_n^h, b_1^t\}$. This operation causes a fusion of two linear chromosomes.

Select two telomeres in one linear chromosome (Figure 2.5h).[1] For example, select telomeres $\{a_1^t\}$ and $\{a_n^n\}$ from linear chromosome $A = (a_1 \dots a_i a_{i+1} \dots a_n)$ (See Figure 2.5h). Then circularize the linear chromosome by connecting its two ends. Two telomeres, $\{a_1^t\}$ and $\{a_n^h\}$, are replaced by by one adjacency, $\{a_1^t, a_n^h\}$. This operation causes a circularization of a linear chromosome.

As mentioned earlier, we do not include a fission that creates a circular intermediate. This choice is based on desired outcomes, not on any notion of mechanism and, in that sense, follows the spirit of the DCJ model itself, since that model's strength is not the verisimilitude of its mechanism, but the simplicity of its formulation and the universality of its set of operations. As we shall see, running our model produces simulated genomes that more closely resemble actual genomes than those produced under a pure DCJ or HP model.

For gene duplication, we uniformly select a position to start duplicating a short segment of chromosomal material and place the new copy to a new position within the genome. We set L_{max} as the maximum number of genes in the duplicated segment and assume that the number of genes in that segment is a uniform random number between 1 and L_{max}. For example, select one segment $a_{i+1} \dots a_{i+L}$ to duplicate and insert the copy between one adjacency $\{b_j^h, b_{j+1}^t\}$. Such an operation duplicates L genes and $L-1$ adjacencies, removes one adjacency, and adds two new adjacencies; thus genes $a_{i+1}, \dots, a_{i+L-1}$ and a_{i+L} are added to the multiset of genes, the adjacency $\{b_j^h, b_{j+1}^t\}$ is removed, and $L+1$ new adjacencies, $\{b_j^h, a_{i+1}^t\}$, $\{a_{i+1}^h, a_{i+2}^t\}$, $\dots$, $\{a_{i+L}^h, b_{j+1}^t\}$, are added.

For gene loss, we uniformly select one gene from the set of all candidate genes and delete it, restricting gene loss to the deletion of a single gene copy at a time, following Lynch [66]. For example, if we delete gene a_i in the chromosome $(\dots a_{i-1} a_i a_{i+1} \dots)$, one copy of a_i is removed from the multiset of genes, while two adjacencies, $\{a^h_{i-1}, a^t_i\}$ and $\{a^h_i, a^t_{i+1}\}$, are replaced by one adjacency, $\{a^h_{i-1}, a^t_{i+1}\}$.

---

[1] Selecting one telomere twice is assimilated to selecting both telomeres of the linear chromosome.

2.2.2 True distance estimation under the new evolutionary model

The problem of estimating the true evolutionary distance is defined as follows:
Input: The original genome G and the final genome F.
Output: An estimate of the actual number of evolutionary events that took place in the evolutionary history to transform G into G^F.

Based on the multisets of genes and of adjacencies and telomeres of G, for any genome G^* of N^* genes and l^* linear chromosomes, we can build the vector $V^* = (NG^*_1, \ldots, NG^*_C, SA^*_1, \ldots, SA^*_C, DA^*, ST^*, DT^*)$, where C is the upper bound for the number of copies of one gene, NG^*_i $(i = 1, \ldots, C)$ is the number of genes with *exactly i* copies in the genome G^*, SA^*_1 $(i = 1, \ldots, C)$ is the number of adjacencies with *exactly i* copies in G^* that also appear in G, DA^* is the number of adjacencies in G^* that do not appear in G, ST^* is the number of telomeres in G^* that also appear in G, and DT^* is the number of telomeres in G^* that do not appear in G. We can write

$$
\begin{aligned}
N^* &= \sum_{i=1}^{C} NG^*_i, \\
N^* &= \sum_{i=1}^{C} SA^*_1 + DA^* + ST^* + DT^* - l^*.
\end{aligned}
$$

Let G^k be the genome obtained from $G = G^0$ by applying k randomly selected evolutionary operations—under our model, the $(i+1)$st evolutionary operation is selected from all possible rearrangements, gene duplications, and gene losses on genome G^i according to the parameters p_d and p_l. We can compute the vector $V_G(G^k) = (NG^k_1, \ldots, NG^k_C, SA^k_1, \ldots, SA^k_C, DA^k, ST^k, DT^k)$ to represent the genome G^k with respect to G.

Now we show that, given G, we can also produce the *estimate* $\widetilde{E}(V_G(G^k)) = (\widetilde{NG^k_1}, \ldots, \widetilde{NG^k_C}, \widetilde{SA^k_1}, \ldots, \widetilde{SA^k_C}, \widetilde{DA^k}, \widetilde{ST^k}, \widetilde{DT^k})$ for the expected vector $E(V_G(G^k))$, for any integer $k > 0$. Our approach for estimating the true evolutionary distance is then to return the integer k that minimizes the 1-norm distance between $\widetilde{E}(V_G(G^k))$ and $V_G(G^F)$.

Estimation of the expected vector after some number of random evolutionary events

Given the original genome G, the complete vector for genome G^k is defined as $V_G(G^k) = (NG^k_1, NG^k_2, \ldots, SA^k_1, SA^k_2, \ldots, DA^k, ST^k, DT^k)$, where NG^k_i is the number of genes with exactly i copies in the genome G^k, SA^k_i (shared adjacencies) is the number of adjacencies with exactly i copies in G^k that also appear in G, DA^k (distinct adjacencies) is the number of adjacencies in G^k that do not appear in G, ST^k (shared telomeres) is the number of telomeres in G^k that also appear in G, and DT^k (distinct telomeres) is the number of telomeres in G^k that do not appear in G.

Assume the original genome G has N genes, where each gene has at most $C = O(1)$ copies, and l linear chromosomes, with $l = O(1)$. We thus ignore items NG^k_i and SA^k_i for $(i > C)$. The initial vector $V_G(G^0)$ is then $(NG^0_1, NG^0_2, \ldots, NG^0_C, SA^0_1, SA^0_2, \ldots, SA^0_C, DA^0, ST^0, DT^0)$, where NG^0_i is the number of genes with exactly i copies, SA^0_i is the number of adjacencies with exactly i copies, $DA^0 = 0$, $ST^0 = 2l$, and $DT^0 = 0$. We now show how to update this vector under rearrangements, gene duplications and gene losses, respectively.

Rearrangement

For rearrangements, we select two adjacencies or telomeres uniformly, with replacement, from the multiset of all adjacencies or telomeres.

Lemma 2.2.1. *Assume all genomes have* $O(1)$ *linear chromosomes, each gene has at most* $C = O(1)$ *copies, and* $V_G(G^k) = (NG_1^k,\ldots, NG_C^k, SA_1^k,\ldots, SA_C^k, DA^k, ST^k, DT^k)$ *represents the current genome* G^k *based on the original genome* G*. For conciseness, write* $N^k = \sum_{i=1}^{C} NG_1^i$ *(the total number of genes) and* $l^k = (ST^k + DT^k)/2$ *(the number of linear chromosomes). Then we can write the expected vector for* G^{k+1} *after one rearrangement operation:* $E(V_G(G^{k+1})) = (NG_1^{k+1},\ldots,NG_C^{k+1},SA_1^{k+1},\ldots,SA_C^{k+1},DA^{k+1}, ST^{k+1},DT^{k+1})$ *where we have*

$$\begin{aligned}
NG_i^{k+1} &= NG_i^k,\ \ i = 1,2,\ldots,C \\
SA_i^{k+1} &= SA_i^k - \frac{2i(SA_i^k - SA_{i+1}^k)}{N^k + l^k} + O(\frac{1}{N^k}),\ \ i = 1,2,\ldots,C-1 \\
SA_C^{k+1} &= SA_C^k - \frac{2C(SA_C^k)}{N^k + l^k} + O(\frac{1}{N^k}), \\
DA^{k+1} &= DA^k + \frac{2(\sum_{i=1}^{C} SA_i^k)}{N^k + l^k} + O(\frac{1}{N^k}), \\
ST^{k+1} &= ST^k - \frac{2ST^k}{N^k + l^k} + O(\frac{1}{N^k}) \\
DT^{k+1} &= DT^k + \frac{2ST^k}{N^k + l^k} + O(\frac{1}{N^k}).
\end{aligned}$$

Proof. In our evolutionary model, each rearrangement operation replaces old adjacencies or telomeres with new ones. Obviously, any rearrangement operation will not change the gene content, so $NG_i^{k+1} (i = 1,2,\ldots,C)$ will be the same.

We first ignore the adjacencies or telomeres in the original genome G created after a rearrangement event. Remember two adjacencies or telomeres are selected with replacement uniformly from the multiset of all adjacencies and telomeres, and the number of all adjacencies or telomeres for genome G^k is $(N^k + l^k)$. For SA_i^k adjacencies with exactly i copies in G^k which also appear in G, the probability that one adjacency is selected once is $\frac{2SA_i^k(N^k+l^k-SA_i^k)}{(N^k+l^k)^2}$, the probability that two different adjacencies are selected is $\frac{SA_i^k(SA_i^k-i)}{(N^k+l^k)^2}$, the probability that same adjacencies at two different sites are selected is $\frac{(i-1)SA_i^k}{(N^k+l^k)^2}$, and the probability that same adjacency at the same site is selected twice is $\frac{SA_i^k}{(N^k+l^k)^2}$. Ignoring the newly created adjacencies or telomeres in the original genome G, with probability $\frac{2SA_i^k(N^k+l^k-SA_i^k)+iSA_i^k}{(N^k+l^k)^2}$, the number of adjacencies with exactly i copies decreases by i, and, with probability $\frac{SA_i^k(SA_i^k-i)}{(N^k+l^k)^2}$, the number of adjacencies with exactly i copies decreases by $2i$. With probability $\frac{2SA_i^k(N^k+l^k-SA_i^k)+SA_i^k}{(N^k+l^k)^2}$, the number of adjacencies with exactly $(i-1)$ copies increases by $(i-1)$, with probability $\frac{SA_i^k(SA_i^k-i)}{(N^k+l^k)^2}$, the number of adjacencies with exactly $(i-1)$ copies decreases by $2(i-1)$, and, with probability $\frac{(i-1)SA_i^k}{(N^k+l^k)^2}$, the number of adjacencies with exactly $(i-2)$ copies increases by $(i-2)$. Considering $i = 1,2,\ldots,C$ and $C = O(1)$, we have

$$\begin{aligned}
SA_i^{k+1} &= SA_i^k - \frac{2i(SA_i^k - SA_{i+1}^k)}{N^k + l^k},\ \ i = 1,2,\ldots,C-1 \\
SA_C^{k+1} &= SA_C^k - \frac{2C(SA_C^k)}{N^k + l^k}, \\
DA^{k+1} &= DA^k + \frac{2(\sum_{i=1}^{C} SA_i^k)}{N^k + l^k}.
\end{aligned}$$

Now, we show that the correction for our ignoring adjacencies or telomeres after a rearrangement event is $O(\frac{1}{N^k})$ for each item. Consider any adjacency (a,b) in G: we might recover it if we select two adjacencies or telomeres containing two genes a and b. Since each gene has at most C copies in the genome, there are at most C^2 pairs of adjacencies or telomeres that may lead to recovery of the adjacency (a,b). So, with probability at most $\frac{C^2}{(N^k+l^k)^2}$, one specific adjacency in G might be created by the rearrangement. Summing up all the $N-l$ adjacencies in G, we see that the correction for ignoring the newly created adjacencies or telomeres in G is $O(\frac{1}{N^k})$.

Similarly, we can get $ST^{k+1} = ST^k - \frac{2ST^k}{N^k+l^k} + O(\frac{1}{N^k})$ and $DT^{k+1} = DT^k + \frac{2ST^k}{N^k+l^k} + O(\frac{1}{N^k})$. $\square$

Gene duplication

For gene duplications, we select uniformly at random an integer between 1 and L_{max} (the maximum number of genes in the duplication segment), then select uniformly at random a position where to start the duplication, then insert the copy at another position selected uniformly at random.

Lemma 2.2.2. *Assume all genomes have $O(1)$ linear chromosomes, each gene has at most $C = O(1)$ copies, no two same genes or adjacencies are within the segment to be duplicated, and $V_G(G^k) = (NG_1^k,\ldots,NG_C^k,SA_1^k,\ldots,SA_C^k,DA^k,ST^k,DT^k)$ represents the current genome G^k based on the original genome G. For conciseness, write $N^k = \sum_{i=1}^C NG_1^i$ (the total number of genes), $l^k = (ST^k + DT^k)/2$ (the number of linear chromosomes) and $L = (L_{max}+1)/2$ (the average number of genes in a duplication segment). Then we approximate the expected vector for G^{k+1} after one duplication operation with $E(V_G(G^{k+1})) = (NG_1^{k+1},\ldots,NG_C^{k+1},SA_1^{k+1},\ldots,SA_C^{k+1},DA^{k+1},ST^{k+1},DT^{k+1})$ where we have*

$$\begin{aligned}
NG_1^{k+1} &= NG_1^k - \frac{L(NG_1^k)}{N^k}, \\
NG_i^{k+1} &= NG_i^k + \frac{iL(NG_{i-1}^k - NG_i^k)}{N^k}, \ i = 2,\ldots,C-1 \\
NG_C^{k+1} &= NG_C^k + \frac{CL(NG_{C-1}^k) + L(NG_C^k)}{N^k}, \\
SA_1^{k+1} &= SA_1^k - \frac{(L-1)SA_1^k}{N^k - l^k} - \frac{SA_1^k - SA_2^k}{N^k + l^k} + O(\frac{1}{N^k}), \\
SA_i^{k+1} &= SA_i^k + \frac{i(L-1)(SA_{i-1}^k - SA_i^k)}{N^k - l^k} - \frac{i(SA_i^k - SA_{i+1}^k)}{N^k + l^k} + O(\frac{1}{N^k}), i = 2,\ldots,C-1 \\
SA_C^{k+1} &= SA_C^k + \frac{C(L-1)SA_{C-1}^k + (L-1)SA_C^k}{N^k - l^k} - \frac{C(SA_C^k)}{N^k + l^k} + O(\frac{1}{N^k}), \\
DA^{k+1} &= DA^k + \frac{(L-1)DA^k}{N^k - l^k} + \frac{\sum_{i=1}^C SA_i^k + DA^k}{N^k + l^k} + O(\frac{1}{N^k}), \\
ST^{k+1} &= ST^k - \frac{ST^k}{N^k + l^k} + O(\frac{1}{N^k}) \\
DT^{k+1} &= DT^k + \frac{ST^k}{N^k + l^k} + O(\frac{1}{N^k}).
\end{aligned}$$

Proof. In our model, we uniformly select a position to start duplicating L genes and transpose it to one new uniformly chosen position within the genome. The expected number of genes or adjacencies with exactly i copies within the duplication segment is $L(NG_i^k)/N^k$ or $(L-1)SA_i^k/(N^k - l^k)$. The probability that the placement of the duplicated segment breaks one adjacency in SA_i^k is $SA_i^k/(N^k + l^k)$.

We again first ignore the adjacencies or telomeres in the original genome G created after a duplication event. Since we assume that no two genes or adjacencies are same within the duplication segment, we have

$$\begin{aligned}
NG_1^{k+1} &= NG_1^k - \frac{L(NG_1^k)}{N^k}, \\
NG_i^{k+1} &= NG_i^k + \frac{iL(NG_{i-1}^k - NG_i^k)}{N^k},\ i = 2,\ldots,C-1 \\
NG_C^{k+1} &= NG_C^k + \frac{CL(NG_{C-1}^k) + L(NG_C^k)}{N^k}, \\
SA_1^{k+1} &= SA_1^k - \frac{(L-1)SA_1^k}{N^k - l^k} - \frac{SA_1^k - SA_2^k}{N^k + l^k}, \\
SA_i^{k+1} &= SA_i^k + \frac{i(L-1)(SA_{i-1}^k - SA_i^k)}{N^k - l^k} - \frac{i(SA_i^k - SA_{i+1}^k)}{N^k + l^k}, i = 2,\ldots,C-1 \\
SA_C^{k+1} &= SA_C^k + \frac{C(L-1)SA_{C-1}^k + (L-1)SA_C^k}{N^k - l^k} - \frac{C(SA_C^k)}{N^k + l^k}. \\
DA^{k+1} &= DA^k + \frac{(L-1)DA^k}{N^k - l^k} + \frac{\sum_{i=1}^C SA_i^k + DA^k}{N^k + l^k}.
\end{aligned}$$

Now, we show that the correction for our ignoring adjacencies or telomeres after a duplication event is $O(\frac{1}{N^k})$ to each item SA_i^{k+1}. Consider any adjacency (a,b) in G: we might recover it if we move gene a next to gene b after the duplication. Since each gene has at most C copies in the genome, there are at most $2LC^2$ possibly duplication operations to recover that adjacency (a,b). There are altogether $O(L(N^k + l^k)^2)$ different duplication operations. So, with probability $O(\frac{1}{(N^k+l^k)^2})$, one specific adjacency in G might be created by the duplication event. Summing up all the $N - l$ adjacencies in G, we see that the correction for ignoring the newly created adjacencies or telomeres in G is $O(\frac{1}{N^k})$.

Similarly, we can get $ST^{k+1} = ST^k - \frac{ST^k}{N^k+l^k} + O(\frac{1}{N^k})$ and $DT^{k+1} = DT^k + \frac{ST^k}{N^k+l^k} + O(\frac{1}{N^k})$. ☐

Gene loss

For gene losses, we uniformly select one gene with at least two copies and delete it.

Lemma 2.2.3. *Assume each gene has at most $C = O(1)$ copies and $V_G(G^k) = (NG_1^k, NG_2^k, \ldots, NG_C^k, SA_1^k, SA_2^k, \ldots, SA_C^k, DA^k, ST^k, DT^k)$ represents the current genome G^k based on the original genome G. For conciseness, write $N^k = \sum_{i=1}^C NG_i^k$ (the total number of genes) and $l^k = (ST^k + DT^k)/2$ (the number of linear chromosomes). Then we can write the expected vector for G^{k+1} after one rearrangement operation as $E(V_G(G^{k+1})) = (NG_1^{k+1}, \ldots, NG_C^{k+1}, SA_1^{k+1}, \ldots, SA_C^{k+1}, DA^{k+1}, ST^{k+1}, DT^{k+1})$, where we have*

$$\begin{aligned}
NG_1^{k+1} &= NG_1^k + \frac{NG_2^k}{N^k - NG_1^k}, \\
NG_i^{k+1} &= NG_i^k - \frac{i(NG_i^k - NG_{i+1}^k)}{N^k - NG_1^k}, i = 2,\ldots,C-1 \\
NG_C^{k+1} &= NG_C^k - \frac{C(NG_C^k)}{N^k - NG_1^k}.
\end{aligned}$$

Proof. In our model of gene loss, one gene with at least two copies is uniformly selected. The number of all possible genes to be deleted is $N^k - NG_1^k$. For NG_i^k $(i > 1)$ genes with exactly i copies in G^k, the probability that one of them is selected and deleted is $\frac{NG_i^k}{N^k - NG_1^k}$. So with probability $\frac{NG_i^k}{N^k - NG_1^k}$, the number of genes with exactly i copies decreases by i and the number of genes with exactly $(i-1)$ copies increases by $(i-1)$. ☐

We ignore the adjacencies or telomeres in the original genome G to be created after one gene loss. For SA_i^k $(i > 2)$ adjacencies with exactly i copies in G^k which also appears in G, it is difficult to compute the number $f_i(del_j)$ of such adjacencies that each single deletion del_j $(j = 1,\dots,N^k - NG_1^k)$ would affect. But we know that each adjacency with exactly i $(i > 2)$ copies must relate to two genes with more than 2 copies, so we have $\sum_{j=1}^{N^k-NG_1^k} f_i(del_j) = 2SA_i^k$. Considering $i = 2,\dots,C$ and $C = O(1)$, we have

$$\begin{aligned} SA_i^{k+1} &= SA_i^k - \frac{2i(SA_i^k - SA_{i+1}^k)}{N^k - NG_1^k},\ i = 2,\dots,C-1 \\ SA_C^{k+1} &= SA_C^k - \frac{2C(SA_C^k)}{N^k - NG_1^k}. \end{aligned}$$

For SA_1^k adjacencies with exactly 1 copy in G^k that also appears in G, it is also difficult to compute the number $f_1(del_j)$ of such adjacencies that each single deletion del_j $(j = N^k - NG_1^k)$ would affect. Assume $DSA_1^k (= \sum_{j=1}^{N^k-NG_1^k} f_1(del_j))$ is the count of genes with at least two copies but related to those adjacencies with exactly 1 copy in G^k that also appear in G. We consider the effect of rearrangements, gene duplications and losses, and we approximate as follows:

$$\begin{aligned} DSA_1^{k+1} &= DSA_1^k + p_r \frac{2(2SA_2^k - DSA_1^k)}{N^k + l^k} \\ &\quad + p_d \left(\frac{2SA_1^k - 2DSA_1^k + 2SA_2^k}{N^k + l^k} - \frac{(L-1)DSA_1^k}{N^k - l^k}\right) \\ &\quad + p_l \frac{2SA_2^k - DSA_1^k(1 + NG_2^k/(N^k - NG_1^k))}{N^k - NG_1^k}, \\ SA_1^{k+1} &= SA_1^k - p_l \frac{DSA_1^k - 2SA_2^k}{N^k - NG_1^k}. \end{aligned}$$

For telomeres, we simply assume $ST^{k+1} = ST^k$ and $DT^{k+1} = DT^k$.

Finally, we also approximate the number of adjacencies RSA^{k+1} that we could thus ignore under rearrangements, gene duplications, and gene losses, and distribute it to the correction of SA_i^k as follows:

$$\begin{aligned} RSA^{k+1} &= (p_r + \frac{1}{2}p_d)(N - l)(N^k/N)^2/(N^k + l^k)^2 \\ SA_i^{k+1} &= SA_i^k + RSA^{k+1} SA_i^k/(N^k - l^k - DA^k), i = 1,\dots,C-1. \end{aligned}$$

Now, given G^0, we estimate $E(V_G(G^k))$ for $k > 0$ by iterating k times the above formulas (using with p_d and p_l); at every step we identify $E(V_G(G^{k-1}))$ with the actual vector $V_G(G^{k-1})$.

Corollary 2.2.4. *The estimated vector* $\widetilde{E}(V_G(G^i)) = (\widetilde{NG_1^i},\dots,\widetilde{NG_C^i}, \widetilde{SA_1^i},\dots,\widetilde{SA_C^i}, \widetilde{DA^i},\widetilde{ST^i},\widetilde{DT^i})$ *for all integers* i $(0 \le i \le k)$ *can be computed in* $O(kC)$ *time.*

2.2.3 Model characteristics

Genome structure prediction

We prove that our new model respects the distinction between eukaryotic and prokaryotic genomes. Note that the following theorems do not deal with the process of chromosome evolution, only with its endpoint.

Theorem 2.2.5. *Let the ancestral genome have one circular chromosome with n genes. After* $O(n)$ *rearrangements events, with probability* $1 - n^{-\Theta(1)}$*, the final genome contains a single circular chromosome or a collection of* $O(\log n)$ *linear chromosomes.*

Proof. We examine the effect of rearrangements on the genome structure. Given the original genome with one circular chromosome, only one of our eight cases can result in a linearization: *select the same adjacency twice* (Figure 2.5f). Once we have only linear chromosomes, two cases can directly result in a change in the number of linear or circular chromosomes: *select the same adjacency twice* (Figure 2.5e) and *select two telomeres* (Figure 2.5h). The probability for selecting the same adjacency twice is $O(1/n)$; that for selecting two telomeres is $O(t^2/n^2)$, where t is the number of telomeres. Every time we select the same adjacency twice, we increase the number of linear chromosomes by 1. Let the indicator variable X_i represent whether or not we select the same adjacency twice at the ith step and write k for the number of evolutionary events. Set $X = \sum_{i=1}^{k} X_i$ and let μ be the expectation of X. The Chernoff bound shows

$$Pr(X > (1+\delta)\mu) < \left(e^{\delta}/(1+\delta)^{1+\delta}\right)^{\mu}$$

In our case, $k = O(n)$, $\mu = O(1)$, $\delta = O(logn)$, so that we get

$$Pr(X > O(\log n)) < n^{-\Theta(1)}$$

Let the indicator variable Y_i represent whether or not we select two telomeres at the ith step. Since $t = 2X$, t is bounded by $O(\log n)$ with probability $1 - n^{-\Theta(1)}$. Thus, with probability $1 - n^{-\Theta(1)}$, we have

$$Pr(Y_i = 1) < O((\log n)^2/n^2).$$

Now set $Y = \sum_{i=1}^{k} Y_i$. We have

$$Pr(Y > 0) \leqslant \sum_{i=1}^{k} Pr(Y_i = 1) < n^{-\Theta(1)}.$$

Overall, then, with probability at least $1 - n^{-\Theta(1)}$, $X < O(\log n)$ and $Y = 0$, which means that the final genome structure has either a collection of $O(\log n)$ linear chromosomes or a single circular chromosome. □

Theorem 2.2.5 tells us that, if the original genomic structure starts from a circular chromosome, most current genomes will contain a single circular chromosome or a collection of linear chromosomes. However, if the initial genome structure was, e.g., a mix of linear and circular chromosomes, would such a structure be stable through evolution? We can characterize all stable structures in our model under some mild conditions.

Theorem 2.2.6. *Let the ancestral genome have n genes and assume that there are positive constants c_1 and α such that each chromosome in the ancestral genome has at least $c_1 n^{\alpha}$ genes. Let c_2 be some constant obeying $c_2 > 2c_1$. After $c_2 n^{1-\alpha} \log n$ rearrangements, with probability $1 - O(n^{-\alpha} \log n)$, the final genome contains either a single circular chromosome or a collection of linear chromosomes.*

Proof. In our evolutionary model, consider the case of selecting two adjacencies or one adjacency and one telomere in two different chromosomes. If one of the two chromosomes is circular, a fusion will merge the circular chromosome into the linear chromosome (Figure 2.5c). If both chromosomes are circular, a fusion will merge the two chromosomes into a single circular chromosome (Figure 2.5d). We use a graph representation, G, for the genome structure, where each circular chromosome is represented by a vertex A_i and all of the linear chromosomes (if any) are represented by a single vertex B . If two adjacencies or one adjacency and one telomere are selected in two different chromosomes, we connect the vertices of these two chromosomes. If we first ignore circularizations of linear chromosomes (Figure 2.5h), then the genome ends up with a single circular chromosome or a collection of linear chromosomes if and only if the corresponding final graph G is connected. We therefore bound the probability that the graph G is not connected after $c_2 n^{1-\alpha}\log n$ rearrangements. If G is not connected, there is at least one bipartition of the vertices into S_1 and S_2 in which no edge has an endpoint in each subset. Assume there are g_1 and g_2 genes in S_1 and S_2, respectively; then $min\{g_1, g_2\} \geqslant c_1 n^{\alpha}$ and $g_1 + g_2 = n$. Since there are at most $\frac{1}{c_1}n^{1-\alpha}$ chromosomes, we can write

$$\begin{aligned} Pr(G \textit{ is not connected}) &< \binom{\binom{g_1}{2}+\binom{g_2}{2}}{c_2 n^{1-\alpha}\log n} \Big/ \binom{\binom{g_1+g_2}{2}}{c_2 n^{1-\alpha}\log n} \\ &< (1-c_1 n^{1-\alpha})^{c_2 n^{1-\alpha}\log n} < O(n^{-2\alpha}) \end{aligned}$$

Let indicator variable X_i represent whether or not we select the same adjacency twice at the ith step (Figure 2.5e,f) and set $X = \sum_{i=1}^{c_2 n^{1-\alpha}\log n} X_i$. We have

$$\begin{aligned} Pr(X_i = 1) &\leqslant 1/n \\ Pr(X > 0) &\leqslant \sum_{i=1}^{c_2 n^{1-\alpha}\log n} Pr(X_i = 1) = O(n^{-\alpha}\log n). \end{aligned}$$

Now we bound the probability of selecting two telomeres in the same linear chromosome (Figure 2.5h), which causes a circularization of this chromosome—the case we deliberately ignored above. For each linear chromosome, there are four possible ways of selecting two corresponding telomeres. Since the number of linear chromosomes l is bounded by $\frac{1}{c_1}n^{1-\alpha}$, there are at most $\frac{4}{c_1}n^{1-\alpha}$ ways to circularize one linear chromosome in all $(n+l)^2$ ways of selecting two adjacencies or telomeres. Again, let indicator variable Y_i represent circularization of one linear chromosome at the ith step and set $Y = \sum_{i=1}^{c_2 n^{1-\alpha}\log n} Y_i$. We have

$$\begin{aligned} Pr(Y > 0) &\leqslant \sum_{i=1}^{c_2 n^{1-\alpha}\log n} Pr(Y_i = 1) \\ &\leqslant 4c_2 \log n / c_1 n^{2\alpha} < O(n^{-2\alpha}\log n) \end{aligned}$$

Thus, with probability $1 - O(n^{-\alpha}\log n)$, we have: G is connected, $X = 0$, and $Y = 0$, so that the final genome contains either a single circular chromosome or a collection of linear chromosomes. □

The restriction on the minimum size of chromosomes in the ancestral genomes is very mild, since the parameter α can be arbitrarily small.

Our model also predicts, for genomes composed of a collection of linear chromosomes, convergence to a certain number of chromosomes, which depends on the total number of genes.

Theorem 2.2.7. *Assume there are n genes and fewer than* $\frac{1+\sqrt{1+4n}}{2}$ *linear chromosomes in the original genome. The number of linear chromosomes increases during rearrangements, converging to* $\frac{1+\sqrt{1+4n}}{2}$.

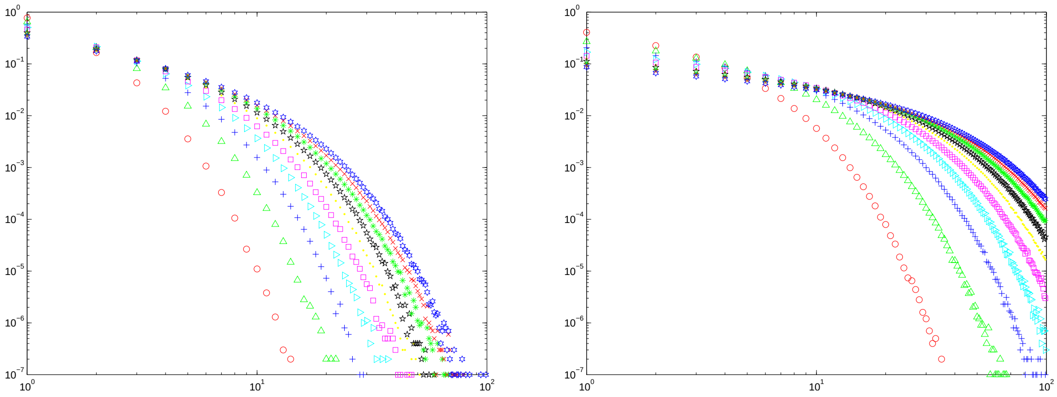
Proof. Assume there are l linear chromosomes in the original genome. In our model, the number of linear chromosomes increases by 1 with probability $\frac{1}{n+l}$ and decreases by 1 with probability $(\frac{l}{n+l})^2$. Since we have $l < \frac{1+\sqrt{1+4n}}{2}$, an increase is more likely. The stable equilibrium follows from the equation $\frac{1}{n+l} = (\frac{l}{n+l})^2$. □

These theorems are not affected by duplications and losses, as long as the latter are reflected in the sizes of chromosomes and the total number of genes.

Sizes of gene families

Of most concern in a duplication and loss model is the distribution of the sizes of the gene families, since that is one of the few aspects of the process that has been observed to obey general laws. Our sole aim in this section is to demonstrate through simulations that our model, which uses the duplication/loss model of Lynch, yields distributions consistent with what Lynch suggested [66].

Our experiments start with a genome with no duplicated genes. This genome is then subjected to a prescribed number k, varying from from 0 to 10 times the number of genes, of evolutionary events chosen according to p_d and p_l to obtain different genomes G^k. We test a large number of different choices of parameters on varying sizes of genomes; as the results are consistent throughout, we report two cases: (a) 1'000 genes with $L = 10$, $p_d = 0.2$, and $p_l = 0.8$; and (b) 10'000 genes with $L = 10$, $p_d = 0.4$, and $p_l = 0.6$. The data in Figure 2.6 summarizes 1'000 runs for each parameter setting. The shape of the distributions of gene family sizes is generally similar to the observations presented by Lynch [66].



(a) 1'000 genes

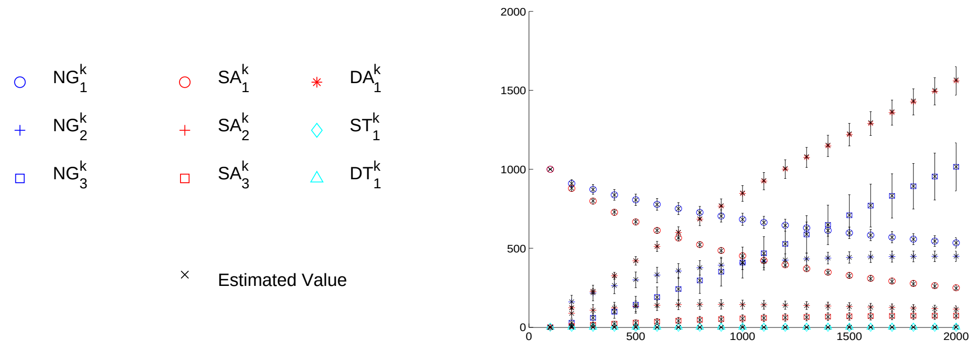
(b) 10'000 genes

Figure 2.6: Probability distribution of the size of gene families, for various numbers of events, increasing from the leftmost ($\#events = \#genes$) to the rightmost ($\#events = 10 \times \#genes$).

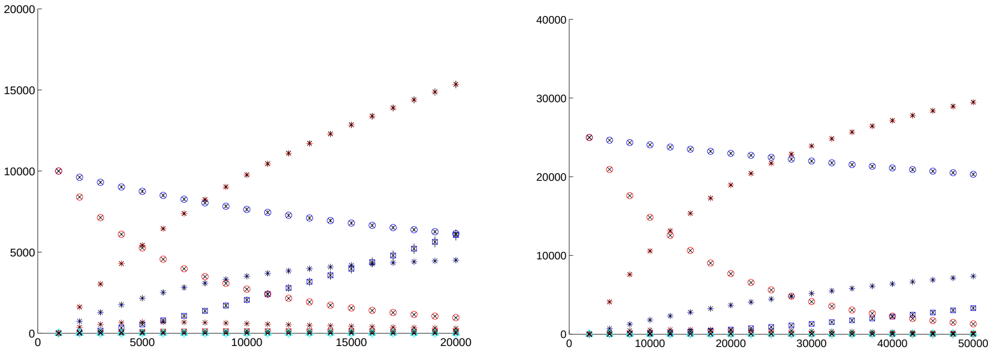
2.2.4 Experimental results

We now present experimental results on the accuracy of our estimation of the expected vector after a given number of random evolutionary events and on the quality of our estimator for the true evolutionary distance (in terms of the actual number of evolutionary events). Our experiments all start with one genome with no duplicated genes and some chosen number of linear and circular chromosomes of various sizes. We first apply some number (usually 10) of duplication events ($L_{max} = 10$ in all cases) to generate the original genome G with some initial duplicated genes. Then this genome is subjected to a prescribed number k of evolutionary events chosen according to p_d and p_l to obtain a final genome G^k. We vary k from 0 to twice the number of genes. We ran tests on any types of initial genomes designed to resemble actual organismal genomes; we tested different choices of parameters on different genomes; and in each case we generated $10,000$ runs to obtain a tight estimate of variance.

We compute the vector representations for all intermediate genomes and then use our method to estimate the evolutionary distance. Due to space limitations, we present results on just three initial genomes: $25,000$ genes and 25 linear chromosomes ($p_d = 0.05, p_l = 0.15$); $10,000$ genes and 5 linear chromosomes ($p_d = 0.1, p_l = 0.2$); and $1,000$ genes and 1 circular chromosome ($p_d = 0.2, p_l = 0.6$).



(a) the nine parameters *(b) 1 circular chromosome, 1'000 genes*



(c) 10 linear chromosomes, 10'000 genes *(d) 25 linear chromosomes, 25'000 genes*

Figure 2.7: The vector values as a function of the actual number of evolutionary events.

The first two examples match large and smaller metazoan genomes, the last matches a small bacterial genome.

Accuracy of the expected vector after k random evolutionary events

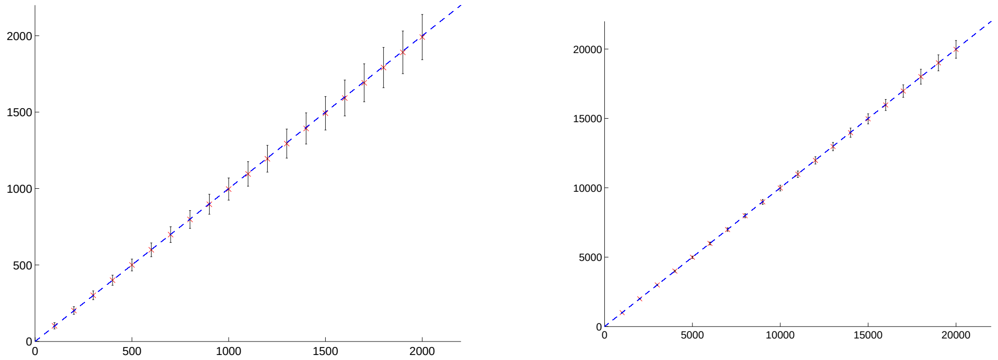
We study the behavior of our estimator $\widetilde{E}(V_G(G^k))$ by comparing its prediction to the sample mean for $V_G(G^k)$, as computed from our $10,000$ trials. In all of our experiments, we find that $\widetilde{E}(V_G(G^k))$ is very close to the sample mean for $V_G(G^k)$. Figure 2.7 shows the values in the vector as a function of the actual number of evolutionary events. SA_3^k and NA_3^k represent the number of adjacencies and genes with at least 3 copies in the original genome G, respectively. The figure shows that our estimation and the sample mean for $V_G(G^k)$ are always very close.

Accuracy of the estimation of the actual number of evolutionary events

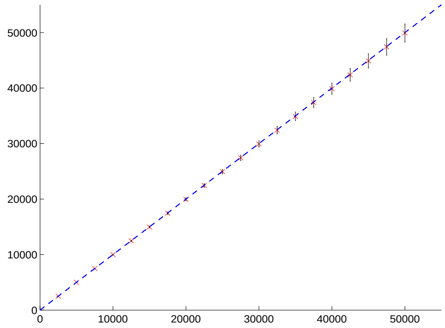
We want to study the accuracy of our estimator for the actual number of evolutionary events; in order to do that, we create simulations with controlled numbers of evolutionary events and set up a threshold for correction in the estimation procedure. Specifically, we vary the actual number of evolutionary events from 0 to twice the number of genes in the original genome and we set 4 times the number of genes as an upper limit on the maximum number of evolutionary events. Thus our estimated number k is chosen to minimize $|\widetilde{E}(V_G(G^k)) - V_G(F)|_1$, the 1-norm distance between $\widetilde{E}(V_G(G^k))$ and $V_G(F)$.

Figure 2.8 shows the mean and standard deviation for the actual number of evolutionary events estimated by our approach. Our approach provides accurate estimates, with very small variance.

We also study the mean absolute difference between the actual number of evolutionary events and our estimator, shown in Figure 2.9.



(a) 1 circular chromosome, 1'000 genes　*(b) 10 linear chromosomes, 10'000 genes*



(c) 25 linear chromosomes, 25'000 genes

Figure 2.8: Mean (×) and standard deviation (vertical bar) for our estimator as a function of the actual number of evolutionary events.

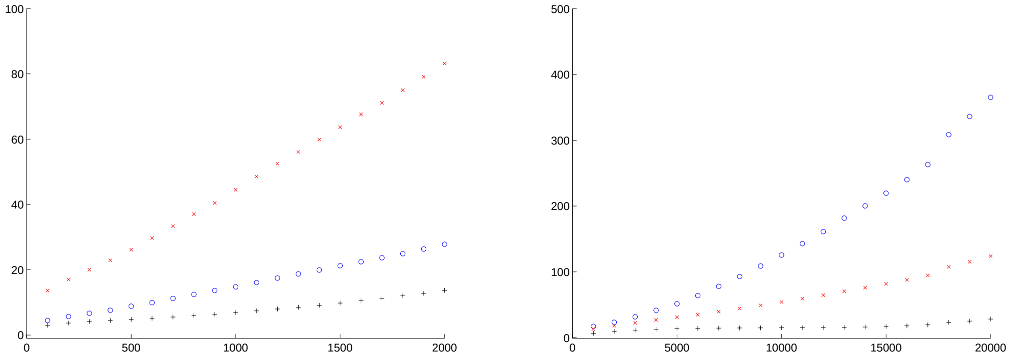
Table 2.2 shows that the estimates are quite accurate up to very large numbers of events. Rearrangements, gene duplications, and gene losses fall under the category of "rare genomic events" (in the terminology of [87]), yet our estimator works well even for numbers that would instead indicate common events.

Robustness to unknown model parameters

Up to now we have fixed p_d and p_l. We now consider the case in which these parameters are unknown—clearly the more common case in practice. We generate $10,000$ cases with randomly parameters p_d and p_l (at 1% resolution, $p_d < 4p_l$) and with actual numbers of evolutionary events varying from 0 to twice the number of genes, setting an upper limit of 4 times the number genes for the maximum number of evolutionary events.

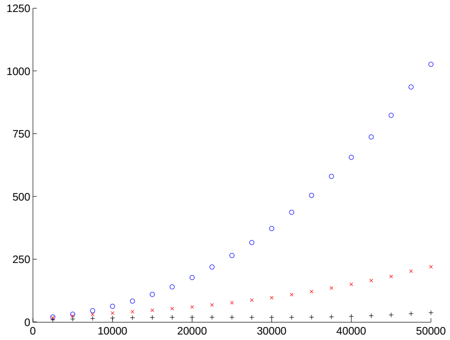
Table 2.2: Relative error of our estimator as a function of the actual number of evolutionary events.

# genes	actual number of evolutionary events					
	# genes $\times 1$			# genes $\times 2$		
	Rearrangements	Duplications	Losses	Rearrangements	Duplications	Losses
1000	7.4 %	3.4 %	7.4 %	6.9 %	3.4 %	6.9 %
10,000	1.7 %	1.4 %	2.7 %	2.6 %	1.4 %	3.1 %
25,000	1.3 %	1.5 %	2.0 %	2.6 %	1.5 %	2.9 %



(a) 1 circular chromosome, 1'000 genes

(b) 10 linear chromosomes, 10'000 genes



(c) 25 linear chromosomes, 25'000 genes

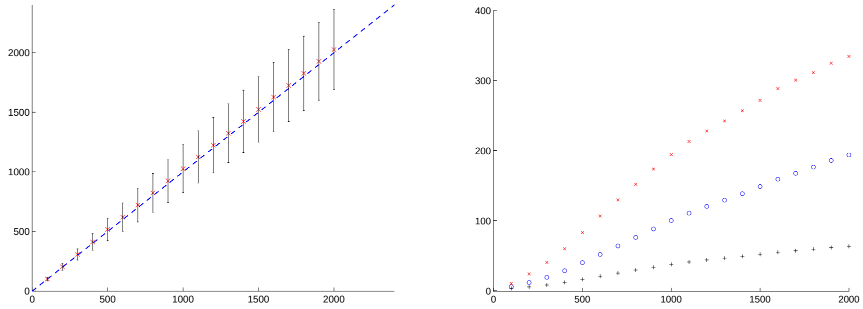
Figure 2.9: The mean absolute difference between the actual number of evolutionary events and our estimation as a function of the actual number of evolutionary events; o: rearrangements, $+$: duplications, $\times$: losses.

Given the original genome, our estimated vector $\widetilde{E}(V_G(G^i))$ is in fact a function of i, p_d, and p_l. We enumerate all possible values for p_d and p_l (at 1% resolution, $p_d < 4p_l$). For each different pair of parameters p_d and p_l, we compute all $\widetilde{E}(V_G(G^i))$ (i from 0 to 4 times the number of genes). Our estimated number k is still chosen to minimize $|\widetilde{E}(V_G(G^k)) - V_G(F)|_1$, the 1-norm distance between $\widetilde{E}(V_G(G^k))$ and $V_G(F)$.

Figure 2.10 shows the comparison of our estimates to the actual number of evolutionary events. Our approach still provides accurate estimates in absence of known values for p_d and p_l and thus is quite robust. The mean absolute difference between the actual number of evolutionary events and our estimator becomes larger, especially when there are few common adjacencies left between the original and final genomes. (The duplications and losses amy also partially cancel each other.)

2.2.5 Discussion

While the mechanism of genome evolution remains unclear, one can nevertheless study different models through simulation and through application to real genomes. Thus, while we make no claims of biological verisimilitude for the operations and constraints within our model, our Theorems 2.2.5 and 2.2.6 show that our model respects the distinction between the organization of most prokaryotic genomes (one circular chromosome) and that of most eukaryotic genomes (multiple linear chromosomes). In contrast, the HP model [40] deals with only linear chromosomes, while the DCJ model [6, 119] (assuming uniform distribution of all possible DCJ operations) predicts that over half of modern genomes consisting of only circular chromosomes will have more than one circular chro-



1 circular chromosome, 1'000 genes

Figure 2.10: Left: mean (×) and standard deviation (vertical bar) of our estimator as a function of the actual number of evolutionary events; right: mean absolute difference between the actual number of different evolutionary events and our estimator (o: rearrangements, +: duplications, ×: losses).

mosome. It is perhaps surprising that a simple modification to the DCJ model (forbidding the least realistic operation) can result in simulated genomes that closely resemble actual genomes—we view this finding as reinforcing the importance of the DCJ model as a basis for future model refinements.

There is evidence about the linearization of circular chromosomes during bacterial evolution [113] and the increase in the number of chromosomes of eukaryotic groups by centric fission [50,51], both of which accord with Theorem 2.2.7. It is interesting to point that Imai *et al.* [52] applied their minimum interaction theory to the genome evolution in eukaryotes to explain the increment in the number of linear chromosomes. Their theory predicts that the highest number of chromosomes in mammals should be 166, while their simulations yield a range of 133–138 for this number [53]. Despite the fact that both models are based on different sets of oversimplified or unrealistic assumptions, the latter range derived by Imai *et al.* is similar to the predictions in our model (as well as the models in [6,40,119], if we assume that the two cuts are uniformly selected) if the number of genes is around 20'000, a fairly typical value for mammals.

Figure 2.6 shows that our model of gene duplications and losses readily generates distributional forms close to the observations presented by Lynch [66]. Different parameters for gene duplications and losses, and the number of evolutionary events, influence the the distributions of gene family sizes: such information can help us improve the estimation of the actual number of evolutionary events as well as infer the parameters for duplications and losses in our model [61,64].

In Section 2.2.4, experimental results on a wide variety of genome structures exemplify the high accuracy and robustness of our estimator. This large gain in accuracy should translate into much better phylogenetic reconstructions as well as more accurate genomic alignments.

According to the analytical results in our model, increasing numbers of rearrangements, gene duplications, and gene losses will linearize circular chromosomes, increase the number of linear chromosomes, and increase the number of genes—i.e., will favor a shift from a prokaryotic architecture to a eukaryotic one. However prokaryotic architectures exist in large numbers today—larger by far than eukaryotic ones. The reason is to be found in population sizes. In a large population, as with most prokaryotic organisms, most alleles are likely to be eliminated by purifying selection, whereas, in a small population, neutral or even deleterious mutations can be fixated more easily. Population sizes decreased dramatically in the transition from prokaryotes to multicellular eukaryotes [66]. Thus many forms of mutant alleles that are able to drift to fixation in multicellular eukaryotes are eliminated by purifying selection in prokaryotes. In a similar way, the fixation of rearrangements, gene duplications, and gene losses (all "rare genomic events" [87]) in prokaryotic species is also more difficult compared to that in eukaryotes. Thus, in our model, prokaryotes tend to have one circular chromosome and a small number of genes, while eukaryotes tend to have multiple linear chromosomes and a large number of genes, in response to a reduction in purifying selection. Our model of gene rearrangement, duplication, and loss is the first to give rise naturally to such a structure; and it does so independently of the choice of parameters, which influence only the tapering rate of the size of gene families.

Chapter 3

Distance-based reconstruction with bootstrapping from whole-genome data

Sankoff and Blanchette [7] introduced the first algorithmic approach to the reconstruction of a phylogenetic tree from whole-genome data, BPAnalysis. The algorithm seeks the tree and internal genomes which together minimize the total number of *breakpoints*—adjacencies present in one genome, but absent in the other. Moret *et al.* [77] reimplemented this approach in their GRAPPA tool and extended it to *inversion distances*—inversions are the best documented of the hypothesized mechanisms of genomic rearrangements. This work focused on unichromosomal genomes; to handle multichromosomal genomes, Bourque and Pevzner [8] proposed MGR, based on GRAPPA's distance computations. Whereas BPAnalysis and GRAPPA search all trees and report the one with the best score (an approach that limits GRAPPA to trees of 15 taxa unless combined with the DCM approach [110], in which case it scales up to 1,000 taxa), MGR uses a heuristic sequential addition method to grow the tree one species at a time. The heuristic approach trades accuracy for scalability, yet MGR does not scale well—in particular, it cannot be used to infer a phylogeny from modern high-resolution data, as even just a few such genomes may require days or weeks of computation. Yet to date MGR (and its more recent derivative MGRA [1]) had remained the only tool available for the analysis of multichromosomal genomic rearrangements. All such parsimony-based approaches must produce good approximations to the NP-hard problem of computing the rearrangement median of three genomes, which limits their scalability [111].

Distance-based methods, in contrast, run in time polynomial in the number and size of genomes—and fast and accurate heuristics exist for those where the scoring function cannot be computed in polynomial time, such as least-squares or minimum evolution methods. Moreover, methods like Neighbor-Joining (NJ) [90] provably return the true tree when given true evolutionary distances. Their speed has long been a major attraction, but the distances that can be computed with sequence data are often far from the true evolutionary distances, particularly on datasets with markedly divergent genomes. Pairwise distances are often computed as edit distances, that is, as minimum-cost distances under the assumed model of evolution. However, even with detailed models, such an edit distance typically underestimates the true distance and that underestimation worsens as the true distance grows. The result is poor trees [73].

The assessment of phylogenies built from whole-genome data has also not been properly addressed to date. The standard method used in sequence-based phylogenetic inference is the bootstrap, but it relies on a large number of homologous characters that can be resampled; yet in the case of whole-genome data, the entire genome is a single character. Alternatives such as the jackknife suffer from the same problem, while likelihood tests cannot be applied in the absence of well established probabilistic models.

In the previous chapter, we have described statistical methods, using exact formulas, to estimate the true evolutionary distance between two genomes under the DCJ model and also under rearrangements, plus gene duplications and losses. In Section 3.1, we show that the high confidence of these estimators can translate into accurate distance-based reconstructions. In Section 3.2, we propose a new approach to the assessment of distance-based phylogenetic inference from whole-genome data.

3.1 Distance-based reconstruction from whole-genome data

3.1.1 Phylogenetic reconstruction and accuracy testing

For reconstruction, we use the distance-based *Neighbor Joining (NJ)* method. Given a matrix of pairwise distances between taxa, NJ reconstructs the phylogeny (including the internal branch lengths) by iteratively joining a closest pair of leaves according to a suitable metric, replacing the two leaves by a "cherry" (the pair of leaves connected to an internal node), computing distances from the cherry to all other leaves, and iterating until only three leaves remain. When the distance matrix is additive, NJ guarantees the reconstruction of the true tree [90].

We study the accuracy of the reconstructed trees and their internal branch lengths through extensive simulations—conducted by generating several trees, simulating evolution on these trees, and using the leaf permutations as inputs to the reconstruction method. The reconstructed trees are compared with the "true" trees to test the accuracy of the method.

We use the Robinson-Foulds (RF) metric [86] to measure the topological accuracy of inferred trees. Every edge e in a leaf-labeled tree defines a bipartition on the leaves: removing e disconnects the tree and thus partitions the set of leaves. If T is the true tree, and T' is the inferred tree, then the false positives are the bipartitions of T' not present in T, and the false negatives are the bipartitions of T not in T'. Divide each count by $n-3$, the number of internal edges in a binary tree on n leaves: the results are the false positive and false negative rates. The RF distance between two binary trees is the average of the number of false negatives and false positives; the RF error rate is the average of the false negative and false positive rates.

The accuracy of branch length estimation is measured by the average branch length error for each inferred tree: $\Sigma|e_i - t_i|/\Sigma t_i$ where e_i and t_i are the edge lengths of edge i in the inferred tree and true tree, respectively, and the summation is over all edges of the trees.

We also study the accuracy of the phylogenetic reconstruction against deviations from the assumed model. We test two scenarios: first by forcing all inversions to be short inversions and second by artificially introducing a fixed number of transpositions. The motivation for selecting short inversions is biological: there is evidence that short inversions are more common than long ones in some prokaryotes and in Drosophila [58, 121]; transpositions have to be artificially introduced because our DCJ model uses two moves to create one transposition.

3.1.2 Experimental design

Our simulation studies follow the standard procedure in phylogenetic reconstruction [42]: we generate model trees under various parameter settings, then use each model tree to produce a number of "true trees" on which we evolve artificial genomes from the root down to the leaves (by performing randomly chosen DCJ operations on the current genome) to obtain datasets of leaf genomes for which we know the complete history. We then reconstruct trees and branch lengths for each dataset by computing a distance matrix using our DCJ-based true distance estimator and then using this matrix as input to NJ. We then compute Robinson-Foulds distances and error rates as well as branch-length errors.

A model tree consists of a rooted tree topology and corresponding branch lengths. The trees are generated by a three-step process. We first generate trees using the birth-death tree generator (from the geiger library) in the software R [85], with a death rate of 0 and various birth rates (data shown below is for a rate of 0.001). The branch lengths in this tree are ultrametric (the root-to-leaf paths all have the same length), so, in the second step, the branch lengths are modified to eliminate the ultrametricity. Choosing a parameter c, for each branch we sample a number s uniformly from the interval $[-c,+c]$ and multiply the original branch length by e^s (we used various values of c; data shown below is for $c = 2$). Finally, we rescale branch lengths to achieve a target diameter D for the model tree (the target diameter is the length of the longest path in the tree). Each branch length now represents the *expected* number of evolutionary operations on that branch. From a single model tree, a set of trees is generated for simulation studies by retaining the same topology and varying the branch lengths by sampling, for each branch in the tree, from a Poisson distribution with a mean equal to that of the corresponding branch length in the model tree.

To test the robustness of the phylogenetic reconstruction when there are deviations from the assumed model, we test it by forcing short inversions and by artificially introducing transpositions. To force short inversions, whenever the random DCJ operation selected is an inversion, the second cut is re-selected (uniformly at random) within a distance s in the same chromosome. We fix $s = 50$ in our experiments. Transpositions are introduced by randomly selecting three cuts, such that the first two cuts are within a chromosome and the third cut is outside the range of the first two cuts. We conducted tests where 20% of the operations in each branch were forced to be (random) transpositions.

All experiments are conducted by varying three main parameters: the number of leaves, the number of genes, and the target diameter. The number of leaves in the trees simulated are 100 and 500, the number of genes are $5,000$ and $10,000$ and the target diameters range from $0.5n$ to $4n$, where n is the number of genes. For each setting of the parameters, 100 model trees are generated and from each model tree 10 datasets are created. The error rates for RF and branch length shown in the next section are averages over these $1,000$ trees.

We also test our reconstruction technique on a real dataset: genomes of 6 species from the Ensembl Mercator/Pecan alignments with 8,380 common markers. We select these genomes for their size, to demonstrate the scalability of our approach, but also because, among vertebrate genomes, they are the best assembled: other vertebrate genomes in the alignment have anywhere from twice to ten times more contigs than the actual chromosomal number of the species.

3.1.3 Experimental results

Simulation studies of the phylogenetic reconstruction

Figure 3.1 shows RF error rates for various trees. The rates for trees with 100 and 500 species, with genomes of size 5,000 and target diameters ranging from 2,500 to 20,000 are shown in Figure 3.1.

The error rates are below 10% in all but the oversaturated cases. The rates for trees of 100 species, with genomes of size 5,000 and 10,000 and diameters varying from half the number of genes to four times that number are shown in Figure 3.1. As expected, error rates are significantly reduced by an increase in the size of the genome—because the larger number of genes reduces the relative error in the estimated distances.

The corresponding average branch-length errors are shown in Figure 3.1.3. Interestingly, the average error in branch length grows more slowly that the RF error rate with increasing evolutionary diameters.

The robustness of the reconstruction method when there are deviations from the evolutionary model is illustrated in Figure 3.3. It shows the RF error rates on trees of 100 and 500 species, with diameter 2000 and with genomes of size 1000. We see that deviations from the model do not affect the accuracy of the reconstruction method.

Overall, these simulations (and many others not shown) confirm that the high precision of our distance estimator makes it possible to reconstruct accurate phylogenies with what is perhaps the simplest of all reconstruction methods, and certainly one of the fastest.

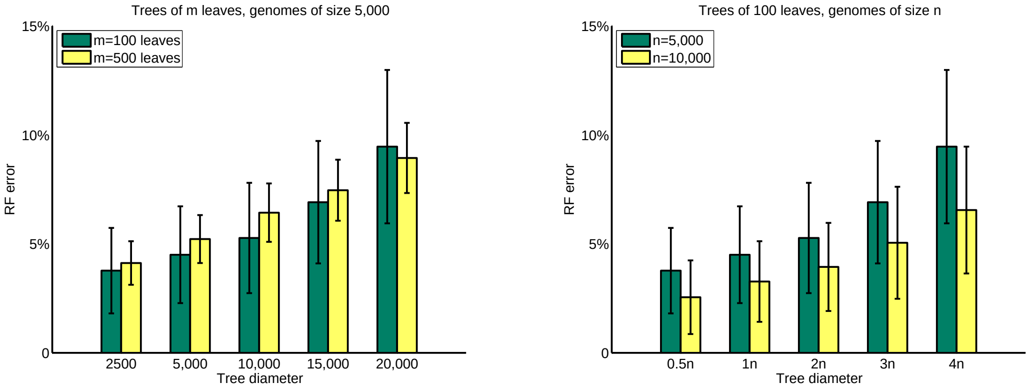


Figure 3.1: RF error rates on trees of 100 and 500 leaves with genomes of size 5,000 (left) and on trees of 100 leaves with genomes of size 5,000 and 10,000 (right).

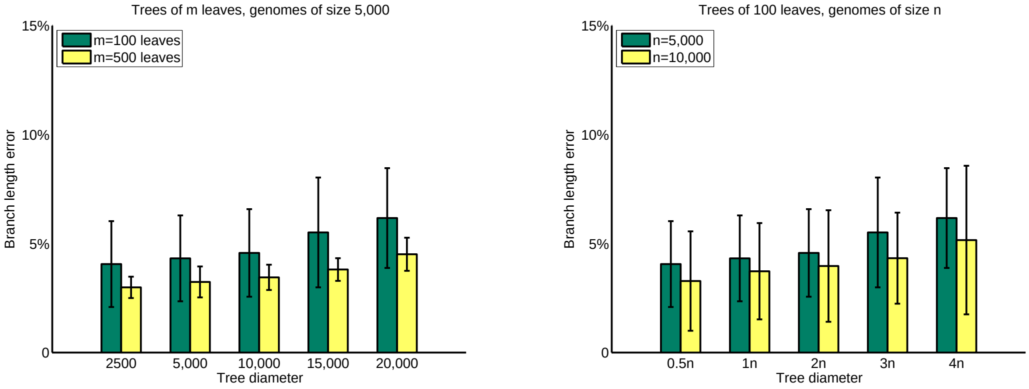


Figure 3.2: Average branch length error on trees of 100 and 500 leaves with genomes of size 5,000 (left) and on trees of 100 leaves with genomes of size 5,000 and 10,000 (right).

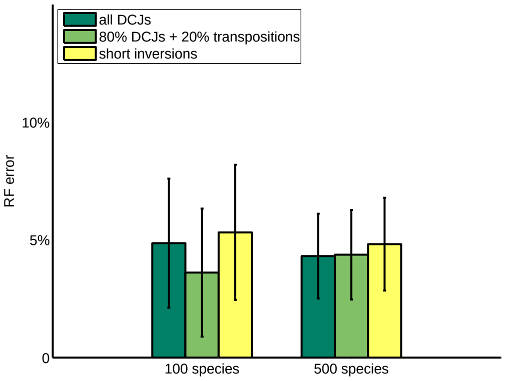


Figure 3.3: RF Error Rates in trees of 100 and 500 species with genomes of sizes 1000

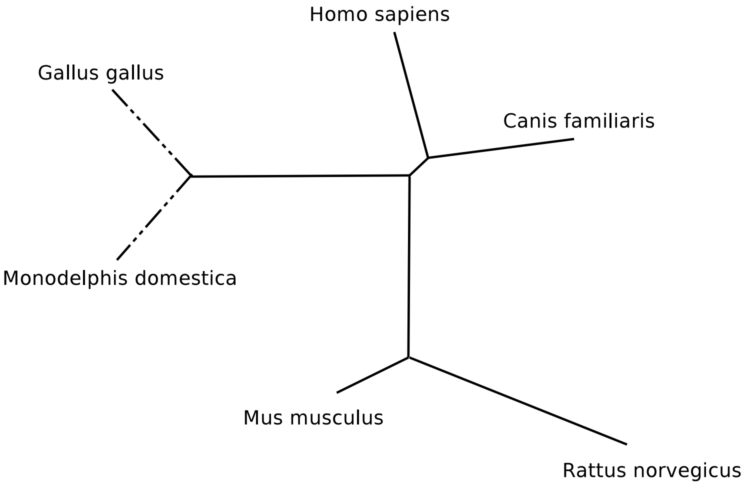


Figure 3.4: Reconstructed phylogeny of man, rat, mouse, opossum, dog, and chicken (dotted edges indicate long branches not shown at scale).

A dataset of high-resolution vertebrate genomes

Figure 3.4 shows the reconstructed phylogeny of 5 mammals and chicken. Building this phylogeny took under a second of computing time on a desktop computer; contrast this very fast computation with the fact that no other tool today can handle this size of genome (over 8,000 syntenic blocks) at all, not even in weeks or months of computation.

The excellent scaling properties of our method means that it is now possible to study the use of whole-genome data in phylogenetic reconstruction, so as to improve our understanding of the evolutionary processes at work, parameterize the model, and eventually make whole-genome data into a source of information for systematics on a par with today's sequence data.

3.1.4 Discussion

We have described a very fast, distance-based, phylogeny reconstruction method for high-resolution whole-genome data. It takes advantage of some of the unique characteristics of whole-genome data, given in terms of syntenic blocks: the absence of duplicates, the equal content among all genomes, and, most importantly, the lack of both homoplasy and saturation in such data, especially when used with high-resolution data. Our simulations demonstrate the accuracy of the reconstruction method and a proof-of-concept application to a small collection of high-resolution vertebrate genomes yields results in line with current findings.

Our methods scale to data of very high resolution (tens of thousands of syntenic blocks) and, because of the very fast running times of distance methods, to large collections of genomes. Therefore, they can be used to study whole-genome data and deepen our understanding of the evolution of the genome, as well as to turn whole-genome data into a genuine source of phylogenetic information.

3.2 Bootstrapping phylogenies

Bootstrapping was introduced by Efron [19] and Felsenstein proposed bootstrapping for phylogeny reconstruction [30]. There are several expositions on these estimation methods at different levels of mathematical detail [20–23, 71], while Soltis and Soltis [98] and Holmes [43] give surveys of bootstrapping in phylogeny reconstruction.

Given n data points $X = \{x_1, \ldots, x_n\}$ and a statistical estimator $E(x_1, \ldots, x_n)$, a *bootstrap replicate* is a fictional dataset $Y = \{y_1^*, \ldots, y_n^*\}$ constructed by sampling with replacement from X. From each such fictional dataset a value of the estimator E can be obtained. The key idea of bootstrapping is that the distribution of values thus obtained closely matches the original distribution of E and can be used to estimate the confidence limits on the estimator. The advantage of the method lies in its applicability to arbitrary and complicated estimators that may be analytically intractable [22, 23].

In phylogeny reconstruction, the standard bootstrap for sequence data [30, 32] samples columns with replacement from a multiple sequence alignment to create a new alignment matrix of identical dimensions. Thus each bootstrap replicate contains the same number of species and the same number of columns per species, but some columns from the original alignment may be duplicated and others omitted. Each column can be viewed as a variable that is drawn from a space of 4^s possible outcomes at each site—assuming nucleic acid sequence data with s species and neglecting insertions, deletions, and ambiguity codes. From each replicate, a tree can be reconstructed using any of the available reconstruction techniques (such as distance-based methods, maximum parsimony, or maximum likelihood). The tree thus obtained from a single bootstrap replicate is a *bootstrap tree*. Many bootstrap trees are generated through repeated sampling and the *bootstrap score* (or *support*) of a branch in the inferred tree is computed as the proportion of the bootstrap trees that contain this branch (viewed as a bipartition of leaves). Soltis and Soltis [98] and Holmes [43] discuss the pros and cons of the approach in phylogeny reconstruction.

A *jackknife* leaves out one observation at a time, thus creating a sample set $X_{(i)} = \{x_1, \ldots x_{i-1}, x_{i+1}, \ldots, x_n\}$. The estimator can be calculated on this new sample. The jackknife often provides a good approximation to the bootstrap, but it fails when the estimator is not smooth; moreover, the number of distinct sample sets is limited to the number of observations. Shao *et al.* [94] found that the generalized "delete-*d*" jackknife works well in practice, even for non-smooth estimators; in this version, d (or some fixed percentage) of the observations are randomly chosen and omitted to create the new sample set. A special case is *parsimony jackknifing* [27] in which an observation is omitted with fixed probability of $1/e$ when creating a new sample set. In such a case, the expected size of the new sample set is $(1 - 1/e)$ times the size of the original set, which corresponds to a modified bootstrapping procedure in which, after sampling, duplicate samples are not added to the new sample set.

No systematic comparison of these methods has been conducted in the context of phylogeny reconstruction. Felsenstein [30] hinted at the equivalence of support values from classical bootstrapping and from 50% jackknifing. Farris *et al.* [28] argued that 50% jackknifing deletes too many characters and does not allow one to maintain a useful relationship between group frequency and support; they advocated the use of parsimony jackknifing. Salamin *et al.* [91] compared bootstrapping and jackknifing in the context of maximum-parsimony reconstruction and reported that bootstrapping and 50%-jackknifing were comparable at confidence levels of 90% and higher. Finally, Mort *et al.* [78] compared bootstrapping with 50% and 33% jackknifing (with and without branch swapping) and reported that all three methods provide similar support values.

A major drawback of phylogenetic reconstruction from whole-genome data has been the lack of any way to assess the robustness of the reconstructed edges. However, the standard bootstrap cannot be applied directly to whole-genome data because the collection of permutations forms a entire character—a single rearrangement or duplication can affect any part of it. In the world of sequence data this is equivalent to an alignment with a single column, albeit one where each character can take any of a huge number of states. Only one approach, jackknifing genes from whole-genome data, has been suggested in the past [95].

3.2.1 Robustness estimation for trees reconstructed from whole-genome data

We design different methods for whole-genome data and devise analogous methods for sequence data (if they do not exist) and vice versa. We study their behavior with both kinds of data with the aim of developing a method for whole-genome data that is as successful as the classic bootstrap is for sequence data. For a method M that operates on sequence data, we denote by M^* the corresponding method for whole-genome data; we use regular font to denote existing methods, bold font to denote the new methods described in this section.

The methods we present here for whole-genome data rely on our distance estimator [60] and so must be used with distance-based reconstruction methods. Our distance estimator computes the estimated true distance between two multichromosomal genomes, based only on the number of shared adjacencies and the number of linear chromosomes in each genome. This limited view of the input data is crucial, as many of the sampling approaches we describe below do not produce valid genome permutations (e.g., because of additional copies of adjacencies), yet still allow us to tally the number of linear chromosomes and of shared adjacencies.

We can view the classical bootstrap for sequence data (hereafter denoted BC) in terms of noise generation. The original multiple sequence alignment gives rise to a distance matrix D. Each replicate dataset created by sampling columns with replacement from the alignment also gives its corresponding matrix B of perturbed pairwise distances. An entry of the replicate matrix corresponding to leaves i and j can thus be written as $B(i,j) = D(i,j) + N(i,j)$ where $N(i,j)$ denotes the perturbation in the distance introduced by the resampling. This noise parameter is hard to characterize exactly, but it leads us to define bootstrapping approaches based on producing increasingly refined estimates of the noise. (In that sense, BP* and **BP** attempt to shape the noise by returning to the underlying evolutionary process of rearrangement or mutation.)

Bootstrapping by adding Gaussian Noise (hereafter denoted **BGN**), adds Gaussian noise of mean 0 to each entry in the distance matrix. The standard deviation is empirically determined to match as well as possible the noise added by BC. Since the noise added during the sampling process in BC is not random, this is a very rough estimate, but a useful comparison point. In the replicate matrices produced by BC, the noise $N(i,j)$ depends on the pairwise distance $D(i,j)$, so the next step is to design a bootstrap method based on pairwise comparisons, hereafter denoted **BPC**. The bootstrap matrix $B(i,j)$ for **BPC** is constructed by calculating the perturbed pairwise distance for each pair: for each pair of sequences i,j, we construct a new pair of sequences i',j' by sampling columns with replacement, where each column has only two characters and set $B(i,j) = D(i',j')$.

An equivalent method **BPC*** can be designed for whole-genome data, albeit with some complications. Since our distance estimator relies on the number of shared adjacencies, a natural choice is to sample adjacencies in the genome. While the evolution of a specific adjacency depends directly on several others, independence can be assumed if we assume that once an adjacency is broken during evolution it is not formed again—an analog of Dollo parsimony, but one that is very likely in whole-genome data due to the enormous state space. For each pair of genomes i,j, we construct two new pairs of genomes. We sample adjacencies from genome i with replacement and use only these adjacencies to compute the distance $D_1(i,j)$ of leaf i to leaf j. (Note that some adjacencies may be overcounted and some omitted.) Then we sample adjacencies from genome j with replacement and use only these adjacencies to compute the distance $D_2(i,j)$ of leaf j to leaf i. Finally, we set $B(i,j) = (D_1(i,j) + D_2(i,j))/2$.

The noise $N(i,j)$ may depend not just on the pairwise distance $D(i,j)$, but also on other distances in the tree, since BC samples columns with replacement for all leaf sequences *at once*. The next step in modeling $N(i,j)$ is thus to sample from all adjacencies (including telomeres). The total number of possible adjacencies (including telomeres) for n syntenic blocks is roughly $2n^2$, but in a given genome there are at most $2n$ adjacencies and each adjacency conflicts with at most $4n$ other adjacencies. Thus, for large genomes, we may assume that adjacencies are independent (if rearrangements happen randomly), just as columns of an alignment are assumed to be independent in BC. We can now mimic closely the sampling procedure of BC in a rearrangement context, producing procedure **BC***. From the list of all possible adjacencies, **BC*** samples with replacement to form a collection of adjacencies; only adjacencies in this collection are then considered in counting the number of shared adjacencies and then estimating the true evolutionary distances between genomes. (Note that some shared adjacencies are counted more than once due to the sampling with replacement.)

We know that classical bootstrapping (BC) is comparable in performance to parsimony jackknifing (which we denote PJ) in the sequence world. PJ is (asymptotically) equivalent to sampling with replacement (as in BC), but without overcounting, that is, when sampling gives a column that has been previously selected, it is not added to the replicate. Thus we can obtain the equivalent of PJ for whole-genome data, call it **PJ***: selected adjacencies are not counted more than once for computing the number of shared adjacencies between leaves. Other versions of jackknifing are similarly easy to design. For instance, a $d\%-$jackknife (dJK) omits $d\%$ of the columns to create a replicate, so, from the set of all adjacencies (in all the leaf genomes) a $d\%$-jackknife (d**JK***) deletes $d\%$ of the adjacencies at random and only the remaining adjacencies are used in estimating the true pairwise distances. In contrast, the previous jackknifing approach for whole-genome data, developed by Shi *et al.* [95], produces replicates by deleting syntenic blocks from the genome: a $d\%$-jackknife, in their method, produces a dataset where $d\%$ of the markers are randomly deleted from all leaf genomes. The authors recommend setting $d = 40$; we call the resulting method JG*. Note that our approach to jackknifing deletes adjacencies instead of markers.

We also design another robustness estimator based on distance perturbation, hereafter denoted **BP***, which permutes each leaf genome through a (randomly chosen) number of random rearrangements, estimates the new pairwise distances, then subtracts from each pairwise estimate the number of rearrangement operations applied to each of the two genomes. The number of operations applied to each genome is chosen from a Gaussian distribution, and so, for each genome, is potentially different. If x operations are applied to leaf i to yield leaf i' and y operations are applied to leaf j to yield leaf j' (where leaves i and j are in the inferred tree and leaves i' and j' in the bootstrap), the expected distance between i' and j' is increased by $(x+y)$ compared to the distance between i and j. To keep the expected pairwise distance after perturbation close to the distance between the corresponding pair of leaves before perturbation, we set the final (perturbed) distance $B(i',j') = D(i',j') - (x+y)$. Thus BP* relies on additivity, a property likely to be respected with whole-genome data due to its huge state space. We can design an equivalent for sequence data: for each sequence, apply some random number of randomly chosen mutations, then estimate all pairwise distances, and finally subtract from that estimate the number of mutations applied in the perturbation step to each of the two sequences—a method we denote **BP**. **BP** is less reliable than **BP***, as it is much more likely that some of the mutations used in the perturbations cancel each other or cancel some of the mutations on the edit path between the two sequences.

In summary, we have designed a bootstrapping procedure, **BC***, that closely mimics the classic bootstrap for phylogenetic reconstruction, BC, and jackknifing procedures, d**JK*** (including, as a special case, **PJ***), that closely mimic the $d\%-$jackknife (and parsimony jackknife PJ). Along the way, we have also designed less refined versions of bootstrapping and their equivalents for sequence data. In our experiments, we use all of these, plus JG*, the marker-based jackknifing approach of Shi *et al.*. A summary of all the methods can be found in table 3.1.

3.2.2 Experimental design

Our simulation studies follow the standard procedure in phylogeny reconstruction (see, e.g., [42]): we generate model trees under various parameter settings, then use each model tree to produce a number of true trees on which we evolve artificial genomes from the root down to the leaves to obtain datasets of leaf genomes for which we know the complete history. Trees are generated by the process

Table 3.1: A summary of all the methods

BGN, **BGN***	Bootstrap by adding Gaussian Noise to the distance matrix.
BPC, **BPC***	Bootstrap by Pairwise Comparisons: for each pair of sequences/genomes, sample columns/adjacencies with replacement to compute distance.
BC, **BC***	Classical Bootstrap: sample columns with replacement to obtain replicate; sample adjacencies with replacement to compute distance matrix.
PJ, **PJ***	Parsimony Jackknifing: choose each column with $1-1/e$ probablity to create replicate; sample adjacencies with replacement and discard duplicates to compute distance matrix.
dJK, **dJK***	d%-JackKnife: Omit d% of columns at random to produce replicate; omit d% of adjacencies at random to compute distance matrix.
BP, **BP***	Bootstrap by Perturbations: apply random mutations/rearrangements to get replicates.
JG*	Jackknife Genes: Marker based jackknifing method for whole-genome data [95].

described in Section 3.1.2. Note that the unit of "length" of an edge is one expected evolutionary operation—mutation or rearrangement.The sequences are evolved by random point mutations under the Kimura 2-parameter (K2P) model (see [109]) using various transition/transversion ratios; the permutations are evolved through double-cut-and-join (DCJ) operations chosen uniformly at random. For sequence data, the distances between leaf sequences are given by the standard distance estimate for the K2P model [109] and the tree is reconstructed with the Neighbor-Joining (NJ) [90]. For rearrangement data, we reconstruct trees by computing a distance matrix using our DCJ-based true distance estimator and then using this matrix as input to both the Neighbor-Joining (NJ) [90] and FastME [17] algorithms.

Experiments are conducted by varying the number of syntenic blocks and the target diameter. We use trees with 100 leaves. Among the many parameter values tested we show the following representative settings: for sequence data, each leaf has 10,000 characters and the tree diameter is 20,000, while, for whole-genome data, we show the results on two sets of parameters, one where each genome has 1,000 markers and the tree diameter is 2,000 and another where each genome has 5,000 markers and the tree diameter is 15,000. For each setting of the parameters, 100 model trees are generated and from each model tree 10 datasets are created; we then average results over the resulting 1,000 trees. For each experiment we produce 100 replicates and thus 100 bootstrap trees from which to compute the bootstrap support of each branch.

A Receiver-Operator-Characteristic (ROC) curve is drawn for every method we investigate. In this plot, a point is a particular bootstrapping test, defined by its *sensitivity* and *specificity*; in the system of coordinates of our figures, a perfect test would yield a point at the upper left-hand corner of the diagram, with 100% sensitivity and 100% specificity. Define E to be the set of edges in the true tree and T_t, for a threshold t, to consist of those edges in the inferred tree that are contained in more than t% of the bootstrap trees. Sensitivity is the proportion of true edges that are also in T_t, $|T_t \cap E|/|E|$, while specificity is the proportion of edges in T_t that are true edges, $|T_t \cap E|/|T_t|$. In our tests we use every fifth value in the range $[0,100]$ as thresholds.

3.2.3 Experimental results

Figure 3.5 shows the ROC curves of the methods for sequence data, for 100 sequences of 10,000 characters each, and a tree diameter of 20,000. The four "reference" methods—50%-jackknifing (50JK), classical bootstrapping (BC), $(1/e)$%-jackknifing (37JK), and parsimony jackknifing (PJ)—are nearly indistinguishable and clearly dominate the others. The analogs of all the other methods developed for whole-genome data (**BP**, **BPC** and **BGN**) are clearly worse than the above four, with **BP** and **BPC** being comparable and the most primitive noise-shaping method, **BGN**, doing the worst.

Figure 3.6 show the ROC curves for whole-genome data for different model conditions. The

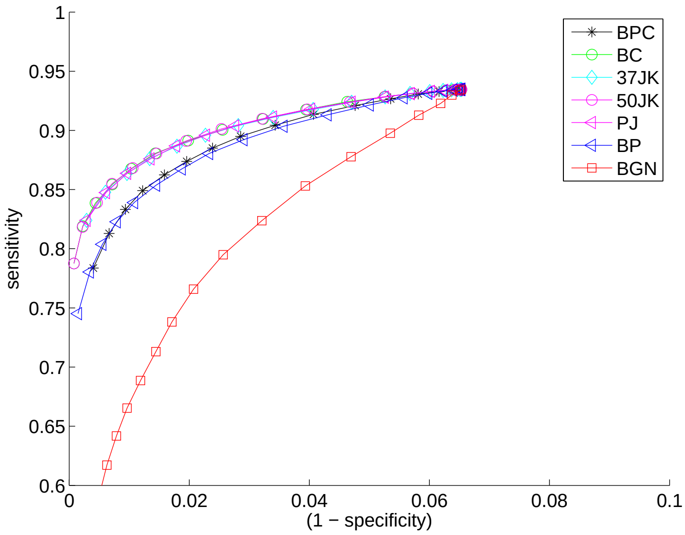


Figure 3.5: *Bootstrapping methods for sequence data*

results follow the same pattern as for sequence data: **BC***, **PJ***, **50JK***, and **37JK*** are nearly indistinguishable and clearly dominate all others. They are followed by **BP*** and **BPC***, which are comparable, while the Gaussian noise approach, **BGN***, again does the worst. JG*, the marker-based jackknifing technique of Shi *et al.*, is better than **BGN***, but trails all other methods. The differences are particularly marked at very high levels of specificity; at 98% specificity, for instance, the top four methods retain nearly 90% sensitivity, but JG* drops to 80%. Very high specificity is the essential characteristic of a good bootstrap method.

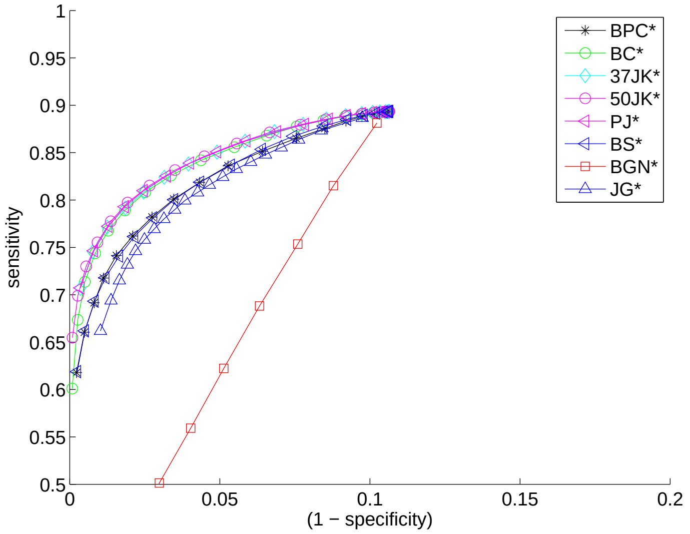
Fig. 3.7 shows the ROC curves for whole-genome data when FastME is used for tree reconstruction instead of NJ. We observe that the relative behavior of the bootstrap methods do not change: **BC***, **PJ***, **50JK***, and **37JK*** perform equally well and dominate other methods. Since the reconstructed trees using FastME are more accurate, the sensitivity of the top four methods at high levels of specificity are even higher compared to the sensitivity attained when NJ was used (Figure 3.6).

A dataset of vertebrate genomes

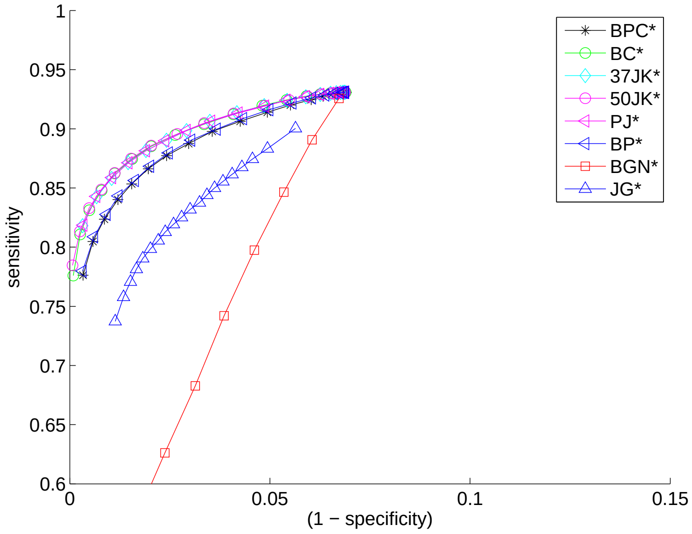
We also test our bootstrapping methods on a real dataset: the genomes of 10 species from the Ensembl Mercator/Pecan alignments with 8,380 common markers. Four of these genomes (horse, chimpanzee, rhesus, and orangutan) are not well assembled: their draft genomes have nearly twice as many contigs as there are chromosomes—but the effect on our adjacency-based distance estimator is minimal, given the large number of markers. Figure 3.8 shows the inferred phylogeny and highlights the two edges with lowest bootstrap support (according to our BC* bootstrapping method). Based on previous studies [2, 11, 49, 69, 79, 118] the edge e_1 is uncertain: some studies place the primates in a clade with rodents, while others place them in a clade with the carnivores. Thus we would expect e_1 to receive the lowest support in the tree. **BC*** does give it the lowest support: 77% for e_1 and 83% for e_2. **BP*** gives low support values for both (49% for e_1 and 44% for e_2), but fails to identify e_1 as the least supported edge, while JG* erroneously gives high support values to both (100% for e_1 and 90% for e_2).

3.2.4 Discussion

Our new approach for whole-genome data, based on the sampling of adjacencies, matches the classical bootstrap and parsimony jackknife approaches and thus provides the first reliable method for

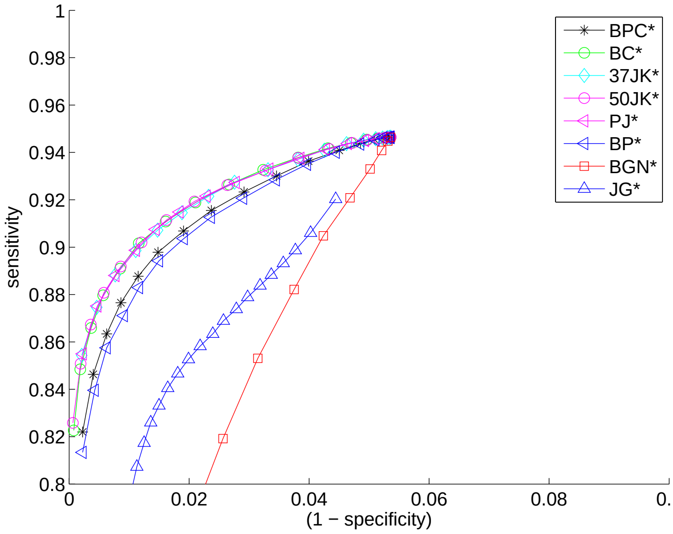


(1,000 markers, diameter 2,000)

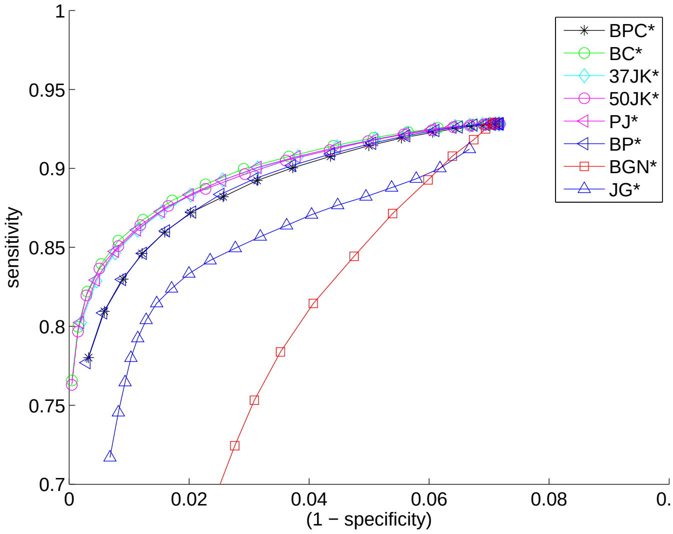


(5,000 markers, diameter 15,000)

Figure 3.6: *Bootstrapping methods for whole-genome data (using NJ)*

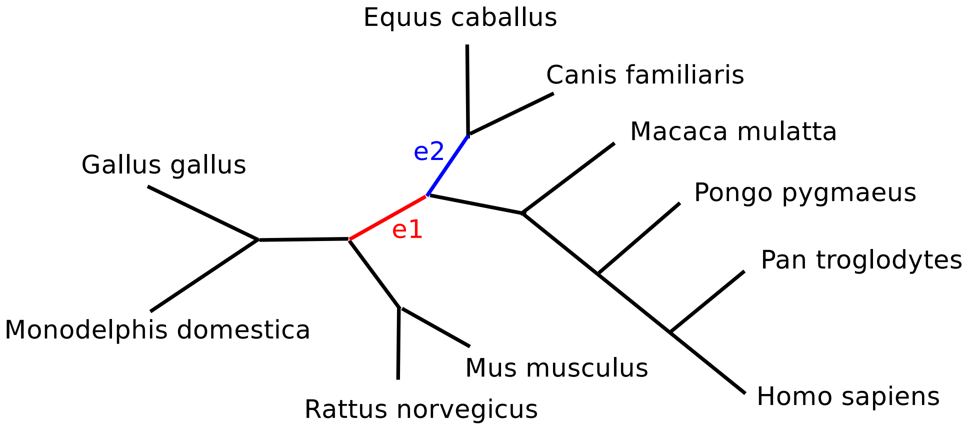


(5,000 markers, diameter 15,000)



(5,000 markers, diameter 20,000)

Figure 3.7: *Bootstrapping methods for whole-genome data (using FastME)*



Equus caballus
Canis familiaris
Macaca mulatta
Gallus gallus
e2
Pongo pygmaeus
e1
Pan troglodytes
Monodelphis domestica
Mus musculus
Homo sapiens
Rattus norvegicus

Figure 3.8: *Inferred phylogeny of 10 vertebrates*

assessing the quality of phylogenetic reconstruction from such data.

In the process of testing various methods, we also confirmed past findings about the superiority of the phylogenetic bootstrap and of the parsimony jackknife [62]. Our results clearly indicate that duplicate samples play no role in the process—parsimony jackknifing works at least as well and occasionally slightly better. Indeed, the best sampling strategy appears to be a random sampling of half of the characters. Given the very high computational cost of the bootstrap, using half the number of characters in sequence-based analyses appears a worthwhile computational shortcut.

Our study focuses on distance-based methods, which reduce the collection of input genomes to a distance matrix. Our basic approach is to equate sampling characters in sequence data with sampling adjacencies in whole-genome data. Any reconstruction method that can handle such data can use this bootstrap procedure. Our reconstruction method is one such method since our distance estimator only counts the number of shared adjacencies between genomes and the number of linear chromosomes in each of them. Possible alternatives for methods (such as Maximum Parsimony) that are unable to handle such data include parsimony jackknifing and direct encoding of adjacencies into sequences. In parsimony jackknifing (**PJ***), each original genome is represented by a set of contiguous regions in the bootstrap; if the reconstruction method can handle such inputs, then this is the best method. Encoding whole-genome data into sequences was proposed many years ago [116] in two different versions (binary encodings and multistate encodings). In such methods, the input is simply a collection of (perfectly) aligned sequences and so the output can be assessed by the standard phylogenetic bootstrap. The early encodings fared poorly in comparison with MP methods (for whole-genome data), but a recent paper [44] suggests that a more complex encoding may overcome these problems.

Chapter 4

Maximum-likelihood reconstruction from whole-genome data

In the previous chapter, we have focused on distance-based methods from whole-genome data. The focus is due to two characteristics of distance-based methods: they are efficient compared to optimization searches such as maximum-likelihood (ML); and it is possible to compute very precise estimates of the true evolutionary distance up to a large value. But distance-based methods still suffer from the problem of *saturation*, that is, if the true distance is too large, the variance of any estimate will be huge, making it impossible to deliver an accurate estimate. In sequence-based phylogenetic analysis, ML approaches has become preferable approaches for distantly related species, although they are much more computationally expensive than distance-based approaches. However, in the last few years, packages such as RAxML [101] have largely overcome computational limitations and allowed reconstructions of large trees (with thousands of taxa) and the use of long sequences (to a hundred thousand characters). It was not until last year that the first successful attempt to use ML reconstruction based on whole-genome data was published [44]; results from this study on bacterial genomes were promising, but somewhat difficult to explain, while the method appeared too time-consuming to handle eukaryotic genomes.

In this chapter, we describe a new approach that resolves these problems and promises to open the way to widespread use of whole-genome data in phylogenetic analysis. Our approach uses a model that includes both rearrangements and duplications and losses; it is robust against common assembly errors; it supports bootstrapping and other standard statistical tests; it returns highly accurate trees in all our tests under a very wide variety of conditions; and it scales as well as approaches based on sequence data. We describe our approach, detail our experimental design, present our results on both simulated and biological data, and discuss our findings on a collection of 68 high-resolution (from 3,000 to over 40,000 genes) eukaryotic genomes from the eGOB database.

4.1 Methods

Our approach encodes the whole-genome data into binary sequences using both gene adjacencies and gene content, then estimates the transition parameters for the resulting binary sequence data, and finally uses sequence-based ML reconstruction to infer the tree. We call our new approach *Maximum Likelihood on Whole-genome Data (MLWD)*.

Encoding genomes into binary sequences

We represent the genome in terms of adjacency information and gene content as follows. Denote the tail of a gene g by g^t and its head by g^h. We write $+g$ to indicate an orientation from tail

Table 4.1: The binary encodings for the two genomes of Figure 1.

	adjacency information						content information			
	$\{a^h,a^h\}$	$\{a^t,b^h\}$	$\{a^t,c^h\}$	$\{b^t,c^t\}$	$\{a^h,d^h\}$	$\{b^t,d^t\}$	a	b	c	d
Genome 1	1	1	1	1	0	0	1	1	1	0
Genome 2	0	1	0	0	1	1	1	1	0	1

to head ($g^t \rightarrow g^h$), $-g$ otherwise ($g^h \rightarrow g^t$). Two consecutive genes a and b can be connected by one *adjacency* of one of the following four types: $\{a^t,b^t\}$, $\{a^h,b^t\}$, $\{a^t,b^h\}$, and $\{a^h,b^h\}$. If gene c lies at one end of a linear chromosome, then we have a corresponding singleton set, $\{c^t\}$ or $\{c^h\}$, called a *telomere*. A *genome* can then be represented as a multiset of adjacencies and telomeres. For example, a toy genome composed of one linear chromosome, $(+a,+b,\text{-}c,+a,+b,\text{-}d,+a)$, and one circular one, $(+e,\text{-}f)$, can be represented by the multiset of adjacencies and telomeres $\{\{a^t\}, \{a^h,b^t\}, \{b^h,c^h\}, \{c^t,a^h\}, \{a^h,b^t\}, \{b^h,d^h\}, \{d^t,a^h\}, \{a^h\}, \{e^h,f^h\}, \{e^t,f^t\}\}$. In the presence of duplicated genes, there is no one-to-one correspondence between genomes and multisets of genes, adjacencies, and telomeres. For example, the genome composed of the linear chromosome $(+a,+b,\text{-}d,+a,+b,\text{-}c,+a)$ and the circular one $(+e,\text{-}f)$, would have the same multisets of adjacencies and telomeres as our toy example.

For data limited to rearrangements (i.e. for genomes with identical gene content), we encode only the adjacency information. For a possible adjacency or telomere, we write 1 (or 0) to indicate its presence (or absence) in a genome. We consider only those adjacencies and telomeres that exist in at least one of the input genomes. If the total number of distinct genes among the input genomes is n, then the total number of distinct adjacencies and telomeres is $\binom{2n+2}{2}$, but the number of adjacencies and telomeres that appear in at least one input genome is typically far smaller—in fact, it is usually linear in n rather than quadratic. For the general model, which includes gene duplications, insertions, and losses in addition to rearrangements, we extend the encoding of adjacencies by also encoding the gene content. For each gene, we write 1 (or 0) to indicate the presence (or absence) of this gene in a genome. For the two toy genomes of Figure 4.1, the resulting binary sequences and their derivation are shown in Table 4.1.

Estimating transition parameters

Since our encodings are binary sequences, the parameters of the model are simply the transition probability from presence (1) to absence (0) and that from absence (0) to presence (1). Let us first look at adjacencies. Every DCJ operation will select two adjacencies (or telomeres) uniformly at random, and (if adjacencies) break them to create two new adjacencies. Each genome has $n+O(1)$ adjacencies and telomeres ($O(1)$ is the number of linear chromosomes in the genome, viewed as a small constant). Thus the transition probability from 1 to 0 at some fixed index in the sequence is $\frac{2}{n+O(1)}$ under one DCJ operation. Since there are up to $\binom{2n+2}{2}$ possible adjacencies and telomeres, the

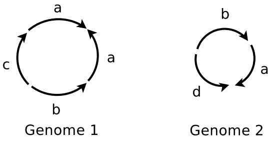


Figure 4.1: Two toy genomes.

transition probability from 0 to 1 is $\frac{2}{2n^2+O(n)}$. Thus the transition from 0 to 1 is roughly $2n$ times less likely than that from 1 to 0. Despite the restrictive assumption that all DCJ operations are equally likely, this result is in line with general opinion about the probability of eventually breaking an ancestral adjacency (high) vs. that of creating a particular adjacency along several lineages (low)—a version of homoplasy for adjacencies.

In the general model, we also have transitions for gene content. Once again, the probability of losing a gene independently along several lineages is high, whereas the probability of gaining the same gene independently along several lineages (the standard homoplasy) is low. However, there is no simple uniformity assumption that would enable us to derive a formula for the respective probabilities—there have been attempts to reconstruct phylogenies based on gene content only [47,97,122], but they were based on a different approach—so we experimented with various values of the ratio between the probability of a transition from 1 to 0 and that of a transition from 0 to 1.

Reconstructing the phylogeny

Once we have the binary sequences encoding the input genomes and have computed the transition parameters, we use the ML reconstruction program RAxML [101] (version 7.2.8 was used to produce the results given in this paper) to build a tree from these sequences. Because RAxML uses a time-reversible model, it estimates the transition parameters directly from the input sequences by computing the base frequencies. In order to set up the $2n$ ratio, we simply add a direct assignment of the two base frequencies in the code.

4.2 Experimental design

We run a series of experiments on simulated datasets in order to evaluate the performance of our approach against a known "ground truth" under a wide variety of settings. We then run our reconstruction algorithm on a dataset of 68 eukaryotic genomes, from unicellular parasites to mammalians, obtained from the *Eukaryotic Gene Order Browser (eGOB)* database [65].

Our simulation studies follow standard practice in phylogenetic reconstruction [42]. We generate model trees under various parameter settings, then use each model tree to evolve an artificial root genome from the root down to the leaves, by performing randomly chosen evolutionary events on the current genome, finally obtaining datasets of leaf genomes for which we know the complete evolutionary history. We then reconstruct trees for each dataset by applying different reconstruction methods and compare the results against the model tree.

Simulating phylogenetic trees

A model tree consists of a rooted tree topology and corresponding branch lengths. The trees are generated by a three-step process. We first generate birth-death trees using the tree generator in the software R [85] (with a birth rate of 0.001 and a death rate of 0), which simulates the development of a model tree under a uniform, time-homogeneous birth-death process. The branch lengths in such trees are ultrametric, so, in the second step, the branch lengths are modified as follows. We choose a parameter c; for each branch we sample a number s uniformly from the interval $[-c, +c]$ and multiply the original branch length by e^s (for the experiments in this paper, we set $c = 2$). Thus, each branch length is multiplied by a possibly different random number. Finally, we rescale all branch lengths to achieve a target diameter D (the length of the longest path, defined as the sum of the edge lengths along that path) for the model tree; each branch length now represents the expected number of evolutionary events on that branch.

Our experiments are conducted by varying three main parameters: the number of taxa , the number of genes, and the target diameter. We used two values for each of the first two parameters: 50 and 100 taxa, and $1,000$ and $5,000$ genes. For the third parameter, the diameter of the tree, we varied it from n to $4n$, where n is the number of genes. For each setting of the parameters, we generated 100 datasets; data presented below are averages over these 100 datasets.

Simulating evolutionary events along branches in the trees

In the rearrangement-only model, all evolutionary events along the branches are DCJ operations. The next event is then chosen uniformly at random among all possible DCJ operations.

In the general model, an event can be a DCJ operation or one of a gene duplication, gene insertion, or gene loss. Thus we sample three parameters for each branch: the probability of occurrence of a gene duplication, p_d, the probability of occurrence of a gene insertion, p_i and the probability of occurrence of a gene loss, p_l. (The probability of occurrence of a DCJ operation is then just $p_r = 1 - p_d - p_i - p_l$.) The next evolutionary event is chosen randomly from the four categories according to these parameters. For gene duplication, we uniformly select a position to start duplicating a short segment of chromosomal material and place the new copy to a new position within the genome. We set $L_{\max}$ as the maximum number of genes in the duplicated segment and assume that the number of genes in that segment is a uniform random number between 1 and $L_{\max}$. In our simulations, we used $L_{\max} = 5$. For gene insertion, we tested two different possible scenarios, one for genomes of prokaryotic type and the other for genomes of eukaryotic type. For the former, we uniformly select one position and insert a new gene; for the latter, we uniformly select one existing gene and mutate it into a new gene. Finally, for gene loss, we uniformly select one gene and delete it.

4.3 Experimental results

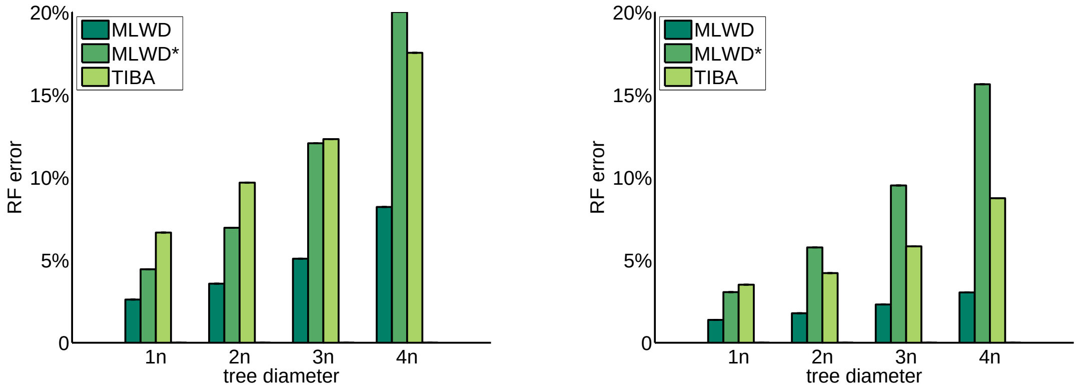
Results for simulations under rearrangements

We compared the accuracy of three different approaches, MLWD, MLWD* and TIBA. MLWD (Maximum Likelihood on Whole-genome Data) is our new approach; MLWD* follows the same procedure as MLWD, but does not use our computation of transition probabilities—instead, it allows RAxML to estimate and set them; finally, TIBA is a fast distance-based tool to reconstruct phylogenies from rearrangement data [63], which combines a pairwise distance estimator [60] and the FastME [16] distance-based reconstruction method. We did not compare with the approach proposed by Hu *et al.* [44], because it is too slow and limited by their character encodings to a maximum of 32 taxa. Figures 4.2 and 4.3 show RF error rates for different approaches; the x axis measures the RF error rates and the y axis indicates the tree diameter.

These simulations show that our MLWD approach can reconstruct much more accurate phylogenies from rearrangement data than the distance-based approach TIBA, in line with experience in sequence-based reconstruction. MLWD also outperforms MLWD*, underlining the importance of estimating and setting the transition parameters before applying the sequence-based ML method.

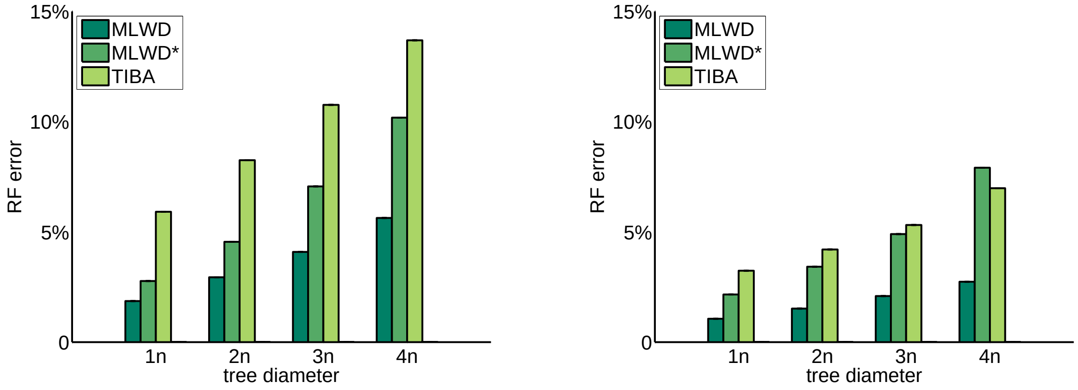
Results for simulations under the general model

Here we generated more complex datasets than for the previous set of experiments. For example, among our simulated eukaryotic genomes, the largest genome has more than 20,000 genes, and the biggest gene family in a single genome has 42 members. Figure 4.4 shows the distributions (averaged over the datasets) of the number of genes and of the size of gene families for datasets of 50 simulated eukaryotic genomes. The encoded sequence of each genome combines both the adjacency and gene content information, which makes it difficult to compute optimal transition probabilities, as discussed



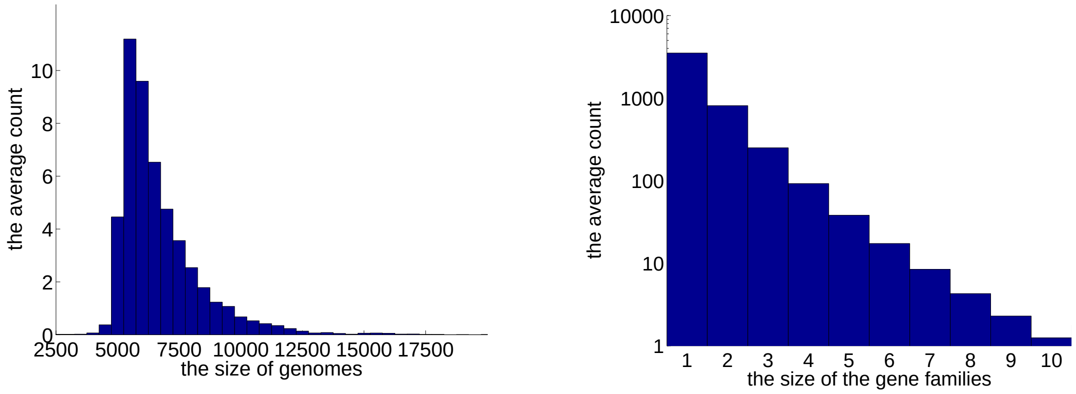
(a) 1'000 genes (b) 5'000 genes

Figure 4.2: RF error rates for different approaches for trees with 50 species, with genomes of 1,000 and 5,000 genes and tree diameters from one to four times the number of genes, under the rearrangement model.



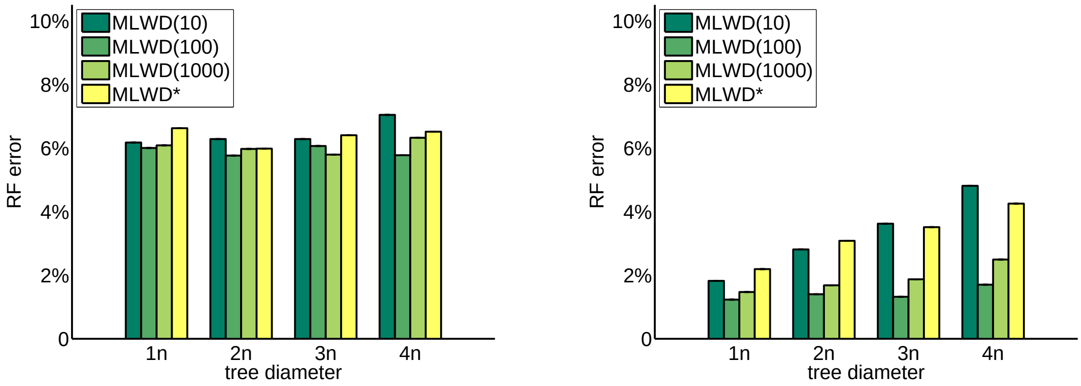
(a) 1'000 genes (b) 5'000 genes

Figure 4.3: RF error rates for different approaches for trees with 100 species, with genomes of 1,000 and 5,000 genes and tree diameters from one to four times the number of genes, under the rearrangement model.



(a) the distribution of the number of genes (b) the distribution of gene family sizes

Figure 4.4: Characteristics of the simulated eukaryotic genomes.



(a) 1'000 genes *(b) 5'000 genes*

Figure 4.5: RF error rates for different approaches for trees with 50 species, with initial genomes of size $1,000$ and $5,000$ and tree diameters from one to four times the number of genes in the initial genome, under the general model of evolution.

in Section 4.1. Thus we set different bias values in our approach and compare them under simulation results. If the transition probability of any gene or adjacency from 0 to 1 in MLWD is set to be m times less than that in the opposite direction, we name it MLWD(m) $(m = 10, 100, 1000)$. Figure 4.5 summarizes the RF error rates. Whereas the best ratio in the rearrangement model was $2n$ (as derived in Section 4.1), the best ratio under the general model is much smaller. This difference can be attributed to the relatively modest change in gene content compared to the change in adjacencies: since we encode presence or absence of a gene, but not the number of copies of the gene, not only rearrangements, but also many duplication and loss events will not alter the encoded gene content.

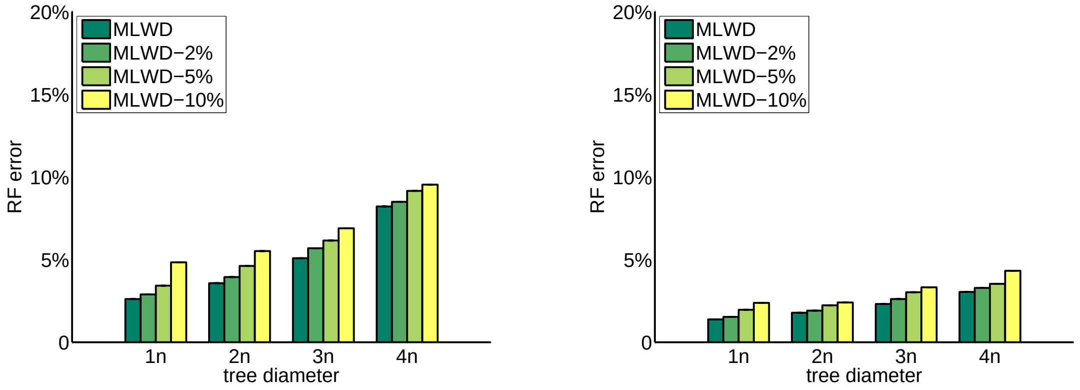
Results for simulated poor assemblies

High-throughput sequencing has made it possible to sequence many genomes, but the finishing steps—producing a good assembly from the sequence data—are time-consuming and may require much additional laboratory work. Thus many sequenced genomes remain broken into a number of contigs, thereby inducing a loss of adjacencies in the source data. In addition, some assemblies may have errors, thereby producing spurious adjacencies and losing others. We designed experiments to test the robustness of our approach in handling genomes with such assembly defects. We introduce artificial breakages in the leaf genomes by "losing" adjacencies, which correspondingly breaks current chromosomes into multiple contigs. For example, MLWD-x% represents the cases of losing x% of adjacencies, that is, x% of the adjacencies are selected uniformly at random and discarded for each genome when the adjacency information for that genome is encoded into binary sequences.

Figure 4.6 shows RF error rates for MLWD on different quality of genome assemblies under the rearrangement model. Our approach is relatively insensitive to the quality of assembly, especially when the tree diameter is large, that is, when it includes highly diverged taxa. Note that this finding was to be expected in view of the good results of our approach using an encoding that, as observed earlier, does not uniquely identify the ordering of the genes along the chromosomes.

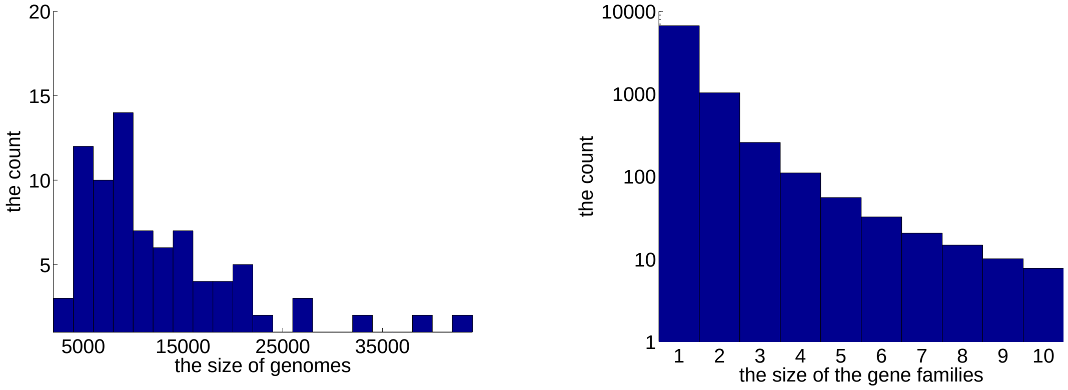
Results for a dataset of high-resolution eukaryotic genomes

Figure 4.8 shows the reconstructed phylogeny of 68 eukaryotic genomes from the eGOB (Eukaryotic Gene Order Browser) database [65]. The database contains the order information of orthologous genes (identified by OrthoMCL [14]) of 74 different eukaryotic species. The total number of different gene markers in eGOB is around $100'000$. We selected 68 genomes for their size (the number of gene



(a) 1’000 genes *(b) 5’000 genes*

Figure 4.6: RF error rates for MLWD on different qualities of genome assemblies, for trees with 50 species, with genomes of size $1,000$ and $5,000$. with tree diameters from one to four times the number of genes, under the rearrangement model.



(a) the distribution of the number of genes *(b) the distribution of gene family sizes*

Figure 4.7: Characteristics of the simulated eukaryotic genomes.

markers) varying from $3k$ to $42k$; the remaining 6 genomes in the database have too few adjacencies (fewer than $3,000$). Figure 4.7 shows the distributions of the genome size and of the size of gene families in those 68 eukaryotic genomes. We encode the adjacency and gene content information of all 68 genomes into 68 binary sequences of length $652'000$. We set the bias ratio to be 100, according to the result of our simulation studies from Section 4.3. Building this phylogeny (using RAxML with fast bootstrapping) took under 3 hours of computing time on a desktop computer. The tree is drawn by the tool iTOL [59]; the internal branches are colored into green, yellow and red, indicating, respectively, strong support (bootstrap value > 90), medium support (bootstrap value between 60 and 90), and weak support (bootstrap value < 60). As shown in Figure 4.8, all major groups in those 68 eukaryotic genomes are correctly identified, with the exception of Amoebozoa. But those incorrect branches with respect to Amoebozoa do receive extremely low bootstrap values (0 and 2), indicating that they are very likely to be wrong. For the phylogeny of Metazoa, the tree is well supported from existing studies [84, 99]. For the phylogeny of model fish species (D. rerio, G. aculeatus, O. latipes, T. rubripes, and T. nigroviridis), two conflicting phylogenies have been published, using different choices of alignment tools and reconstruction methods for sequence data [80]. Our result supports the second phylogeny, which is considered as the correct one by the authors in their discussion [80]. For the phylogeny of Fungi, our results agree with most branches for common species in recent studies [35, 114]. It is worth mentioning that among three Chytridiomycota species C. cinereus, P.

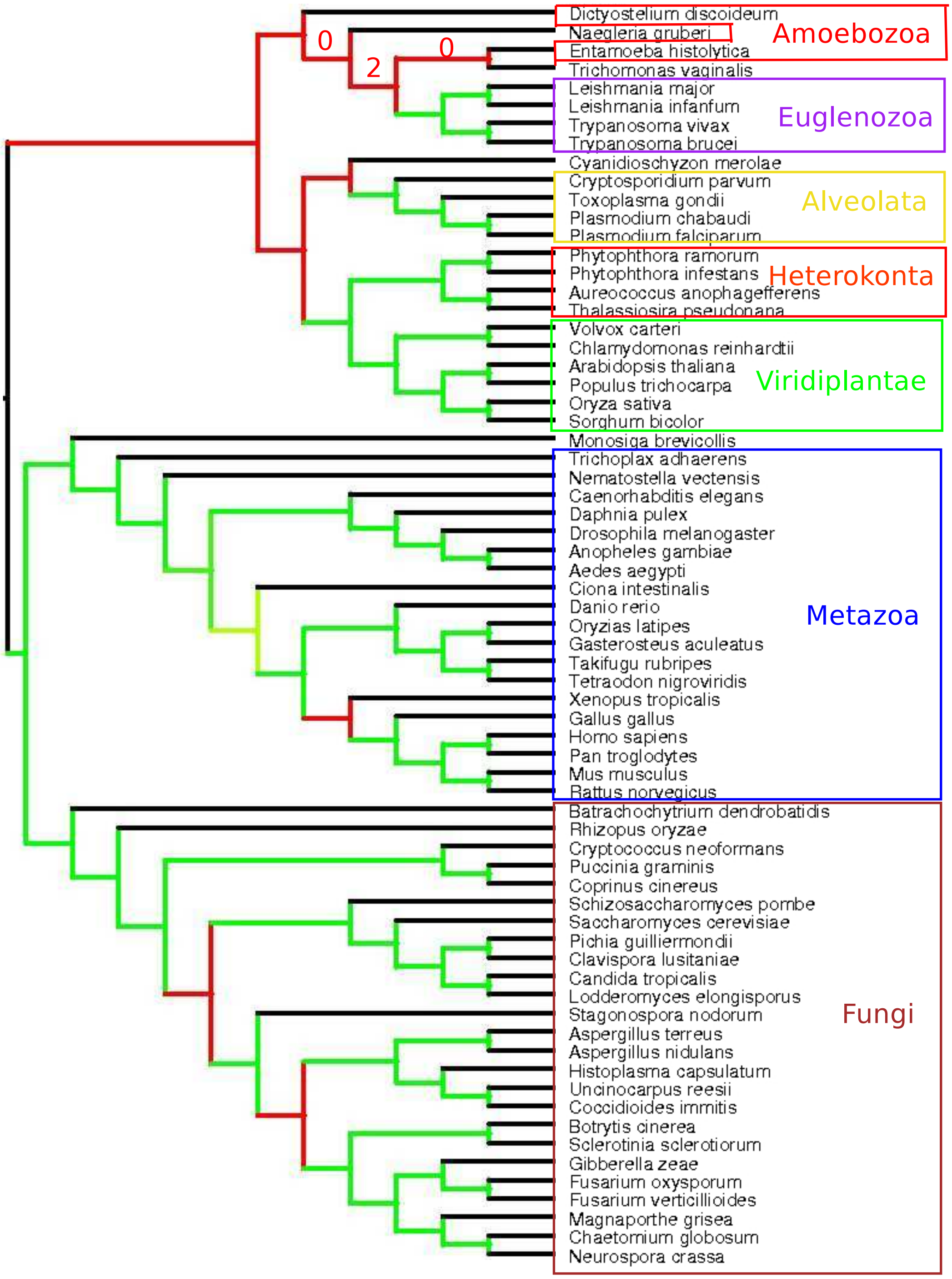

0
2
0
Dictyostelium discoideum
Naegleria gruberi
Entamoeba histolytica
Trichomonas vaginalis
Amoebozoa
Leishmania major
Leishmania infanfum
Trypanosoma vivax
Trypanosoma brucei
Euglenozoa
Cyanidioschyzon merolae
Cryptosporidium parvum
Toxoplasma gondii
Plasmodium chabaudi
Plasmodium falciparum
Alveolata
Phytophthora ramorum
Phytophthora infestans
Aureococcus anophagefferens
Thalassiosira pseudonana
Heterokonta
Volvox carteri
Chlamydomonas reinhardtii
Arabidopsis thaliana
Populus trichocarpa
Oryza sativa
Sorghum bicolor
Viridiplantae
Monosiga brevicollis
Trichoplax adhaerens
Nematostella vectensis
Caenorhabditis elegans
Daphnia pulex
Drosophila melanogaster
Anopheles gambiae
Aedes aegypti
Ciona intestinalis
Danio rerio
Oryzias latipes
Gasterosteus aculeatus
Takifugu rubripes
Tetraodon nigroviridis
Xenopus tropicalis
Gallus gallus
Homo sapiens
Pan troglodytes
Mus musculus
Rattus norvegicus
Metazoa
Batrachochytrium dendrobatidis
Rhizopus oryzae
Cryptococcus neoformans
Puccinia graminis
Coprinus cinereus
Schizosaccharomyces pombe
Saccharomyces cerevisiae
Pichia guilliermondii
Clavispora lusitaniae
Candida tropicalis
Lodderomyces elongisporus
Stagonospora nodorum
Aspergillus terreus
Aspergillus nidulans
Histoplasma capsulatum
Uncinocarpus reesii
Coccidioides immitis
Botrytis cinerea
Sclerotinia sclerotiorum
Gibberella zeae
Fusarium oxysporum
Fusarium verticillioides
Magnaporthe grisea
Chaetomium globosum
Neurospora crassa
Fungi

Figure 4.8: The reconstructed phylogeny of 68 eukaryotic genomes

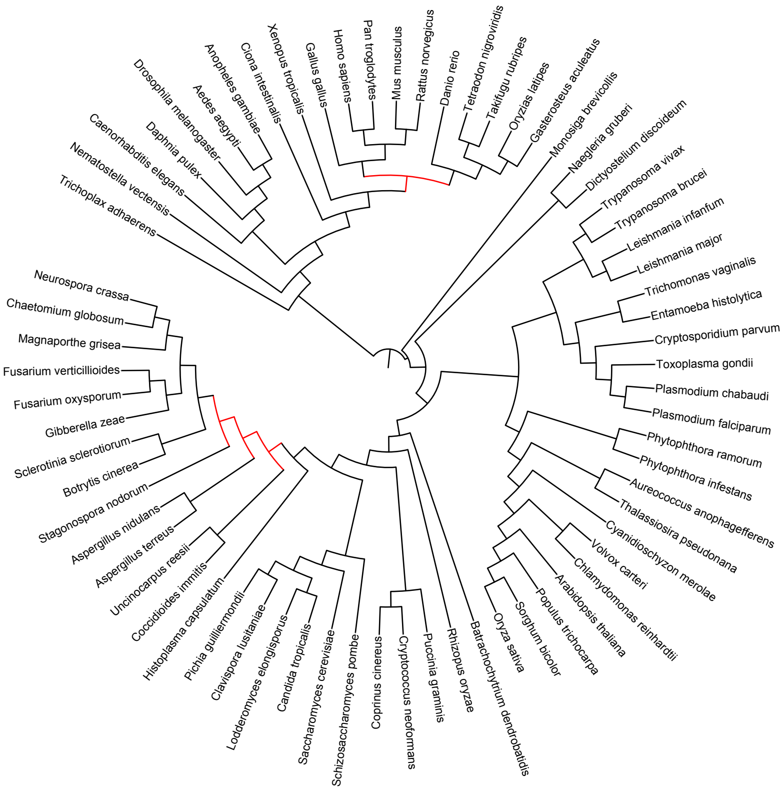

Neurospora crassa
Chaetomium globosum
Magnaporthe grisea
Fusarium verticillioides
Fusarium oxysporum
Gibberella zeae
Sclerotinia sclerotiorum
Botrytis cinerea
Stagonospora nodorum
Aspergillus nidulans
Aspergillus terreus
Uncinocarpus reesii
Coccidioides immitis
Histoplasma capsulatum
Pichia guilliermondii
Clavispora lusitaniae
Lodderomyces elongisporus
Candida tropicalis
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Coprinus cinereus
Cryptococcus neoformans
Puccinia graminis
Rhizopus oryzae
Batrachochytrium dendrobatidis
Oryza sativa
Sorghum bicolor
Populus trichocarpa
Arabidopsis thaliana
Chlamydomonas reinhardtii
Volvox carteri
Cyanidioschyzon merolae
Thalassiosira pseudonana
Aureococcus anophagefferens
Phytophthora infestans
Phytophthora ramorum
Plasmodium falciparum
Plasmodium chabaudi
Toxoplasma gondii
Cryptosporidium parvum
Entamoeba histolytica
Trichomonas vaginalis
Leishmania major
Leishmania infanfum
Trypanosoma brucei
Trypanosoma vivax
Dictyostelium discoideum
Naegleria gruberi
Monosiga brevicollis
Gasterosteus aculeatus
Oryzias latipes
Takifugu rubripes
Tetraodon nigroviridis
Danio rerio
Rattus norvegicus
Mus musculus
Pan troglodytes
Homo sapiens
Gallus gallus
Xenopus tropicalis
Ciona intestinalis
Anopheles gambiae
Aedes aegypti
Drosophila melanogaster
Daphnia pulex
Caenorhabditis elegans
Nematostella vectensis
Trichoplax adhaerens

Figure 4.9: The phylogeny of 68 genomes using only gene contents.

gramnis, and C. neoformans, our phylogeny shows that C. cinereus and P. gramnis are more closely related, which conflicts with the placement of C. cinereus and C. neoformans as sister taxa, but with very low support value (bootstrapping score 35) [114]. C. merolae, a primitive red algae, has been the topic of a longrunning debate over its phylogenetic position [81]. Our result suggests that C. merolae is closer to Alveolata than to Viridiplantae, in agreement with a recent finding obtained by sequencing and comparing expressed sequence tags from different genomes [10].

Finally, in order to explore the relationship between gene content and gene order, we ran MLWD$*$ on the 68 eukaryotic genomes using only adjacency information as well as using only content information. The tree reconstructed from adjacency information only is poor, with even major clades getting mixed—an unsurprising result in view of the huge variation in gene content among these 68 genomes. The tree reconstructed from gene-content information only correctly identifies all major groups except Amoebozoa; however, it suffers from some major discrepancies with our current understanding of several clades, highlighted as red branches in Figure 4.9. For example, X. tropicalis is thought to be closer to mammals than to fishes [26]. H. capsulatum, U. reesii, and C. immitis are considered to be in the same order (Onygenales); together with A. nidulans and A. terreus they are considered to be in the same class (Eurotiomycetes), but S. nodorum is thought to belong to a different class (Dothideomycetes) [114].

4.4 Discussion

In spite of many compelling reasons for using whole-genome data in phylogenetic reconstruction, practice to date has continued to use selected sequences of moderate length using nucleotide-, aminoacid-, or codon-level models. Such models are of course much simpler and much better studied than models for the evolution of genomic architecture. Mostly though, it is the lack of suitable tools that has prevented more widespread use of whole-genome data: previous tools all suffered from serious problems, usually combinations of oversimplified models, poor accuracy, poor scaling, lack of robustness against errors in the data, and lack of any bootstrapping or other statistical assessment procedures.

The approach we presented is the first to overcome all of these difficulties: it uses a fairly general model of genomic evolution (rearrangements plus duplications, insertions, and losses of genomic regions), is very accurate, scales as well as sequence-based approaches, is quite robust against typical assembly errors and omissions of genes, and supports standard bootstrapping methods. Our analysis of a 68-taxon collection of eukaryotic genomes, ranging from parasitic unicellular organisms with simple genomes to mammals and from around 3,000 genes to over 40,000 genes, could not have been conducted, regardless of computational resources, with any other tools without accepting severe compromises in the data (e.g., equalizing gene content) or the quality of the analysis (by using a distance-based reconstruction method). Our analysis also helps make the case for phylogenetic reconstruction based on whole-genome data. We did not need to choose particular regions of genomes nor to process the data from the eGOB database in any manner; in particular, we did not need to perform a multiple sequence alignment. We were able to run a complete analysis on a “Tree of Life” of all main branches of the Eukaryota, with very divergent genomes (and hence very large pairwise distances), without taking any special precautions and without preinterpreting the data (and thus possibly biasing the output). We could do all of this in a few hours on a desktop machine—in spite of the very long sequences produced by our encoding. We could run the identical software on a collection of organellar genomes or of bacterial genomes with equal success (and in much less time).

Chapter 5

Conclusion and discussion

The rapid accumulation of whole-genome data has renewed interest in the study of the evolution of genomic architecture, under such events as rearrangements, duplications, losses. Comparative genomics, evolutionary biology, and cancer research all require tools to elucidate the mechanisms, history, and consequences of those evolutionary events, while phylogenetics could use whole-genome data to enhance its picture of the Tree of Life. Current approaches in the area of phylogenetic analysis are limited to very small collections of closely related genomes using low-resolution data (typically a few hundred syntenic blocks); moreover, these approaches typically do not include duplication and loss events.

There are several improvements in phylogenetic reconstruction presented in this dissertation, each eliminating one or more of these problems that have prevented widespead use of whole-genome data.

Models and distance estimation on whole-genome evolution

We present a method to estimate the true evolutionary distance between two genomes under the ‘double-cut-and-join’ (DCJ) model of genome rearrangements, a commonly used model under which a single multichromosomal operation accounts for all genomic rearrangement events: inversion, transposition, translocation, block interchange and chromosomal fusion and fission. The estimator relies on a simple structural characterization of a genome pair and is both analytically and computationally tractable. To handle rearrangements, gene duplications and losses, we propose a new evolutionary model and the corresponding method for estimating true evolutionary distance. Our model, inspired from the DCJ model, is simple and the first to respect the structural dichotomy in genomic organization (1-2 circular chromosomes vs. several larger linear chromosomes) between most prokaryotes and most eukaryotes. We give the corresponding estimate of genomic pairwise distances under the new evolutionary model, which should translate into much better phylogenetic reconstructions as well as more accurate genomic alignments.

Distance-based reconstruction with bootstrapping from whole-genome data

We present a novel approach to the assessment of distance-based phylogenetic inference from whole-genome data. Our approach restates the main characteristics of the jacknife and bootstrap in terms of noise shaping, itself a longstanding approach to robustness assessment in engineering. For each feature of our method, we give an equivalent feature in the sequence-based framework and present the results of extensive experimental testing, in both sequence-based and genome-based frameworks, demonstrating that our bootstrapping approach for whole-genome data is on par with the classic phylogenetic bootstrap used in sequence-based reconstruction. We test our approach on a small dataset of mammalian genomes, verifying that the support values match current thinking about the respective branches. Our method is the first to provide a standard of assessment to match that of the classic phylogenetic bootstrap for aligned sequences. Thus our assessment method makes it possible to conduct phylogenetic analyses on whole genomes with the same degree of confidence as for analyses on aligned sequences.

Maximum-likelihood reconstruction from whole-genome data

We present a maximum-likelihood approach to phylogenetic analysis that takes into account genome rearrangements as well as duplications, insertions, and losses. Our approach is robust against common assembly errors; it supports bootstrapping and other standard statistical tests; it returns highly accurate trees in all our tests under a very wide variety of conditions; and it scales as well as approaches based on sequence data. The results of extensive testing on simulated data show that our approach returns very accurate results very quickly. In particular, we analyze a 68-taxon collection of eukaryotic genomes, ranging from parasitic unicellular organisms with simple genomes to mammals and from around 3000 genes to over 40000 genes; the analysis, including bootstrapping, takes just 3 hours on a desktop system and returns a tree in agreement with all well supported branches, while also suggesting resolutions for some disputed placements.

Naturally, much work remains to be done. In particular, given the complexity of genomic architecture, current evolutionary models (such as the one we used) are too simple, although even at that level, we need to elucidate simple parameters, such as the ratio of the transition probabilities between loss and gain of a given gene. Using different transition probabilities for adjacencies and for content, by running a compartmentalized analysis, should prove beneficial on larger datasets. Larger issues of data preparation loom. For instance, moving from an assembled genome to the type of data we used (the ordering of genes along each chromosome) continues to require manual intervention—gene-finding, or syntenic block decomposition, are too complex for fully automated procedures. The interplay between data resolution (number of markers) and quality of the resulting tree remains to be explored. Indeed, most of the methodological questions that the phylogenetic community has been studying for the last several dozen years in the context of sequence-based reconstruction also arise, in suitably modified terms, in the context of whole-genome data—but in the latter case, almost all are unaddressed.

It is interesting to study whole-genome evolution by exploring both large-scale changes (e.g., rearrangements, duplications and losses) and local changes (e.g. point mutations and small insertions and deletions). Such a study includes three major steps. The first step is the reconstruction of the phylogeny; our current work provides a possible solution. The second step is ancestral reconstruction. Here we envision propagating the adjacency information from leaves to ancestors and inferring partial contigs with bootstrap scores for ancestor genomes. The third step is the characterization of large-scale and local changes. We envision setting up correspondences between large segments of genomes by selecting sets of non-conflicting and well-supported contigs. Such correspondences will enable us to derive the parameters in our evolutionary model, to identify the large-scale changes, and to clarify local changes within large segments in a divide-and-conquer strategy. Our goal is to provide a modern scientific view of evolution at different resolutions, a view that reveals mechanisms of genomic instability and thus facilitates biomedical research.

Human cancers are associated with the somatic acquisition of a series of DNA sequence variants and mutations. Such variants and mutations fall under the categories of large-scale changes (whole-genome data) and local changes (sequence data) discussed in this work. Therefore techniques developed to characterize such changes could be applied to the study of cancer genomes. Cancer genomes are thought to accumulate chromosomal rearrangements in a relatively short period of time; analyzing genomic data for such cell lines requires fast and reliable tools to handle different time scales. Cancer cells are usually associated with copy number variations across the whole genome; thus a development model for cancer genomes needs to account for massive gene duplications and losses. These various characteristics are all part of the modeling effort we have pursued, and this work provides a promising avenue of exploration to model and search for mechanisms that underlie the development of cancer genomes.

Bibliography

[1] M.A. Alekseyev and P.A. Pevzner. Breakpoint graphs and ancestral genome reconstructions. *Genome Research*, 19(5):943–957, 2009.

[2] H. Amrine-Madsen, K.-P. Koepfli, R.K. Wayne, and M.S. Springer. A new phylogenetic marker, apolipoprotein b, provides compelling evidence for eutherian relationships. *Molecular Phylogenetics and Evolution*, 28(2):225–240, 2003.

[3] M. Anisimova and O. Gascuel. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst. Biol.*, 55(4):539–552, 2006.

[4] D.A. Bader, B.M.E. Moret, and M. Yan. A fast linear-time algorithm for inversion distance with an experimental comparison. *J. Comput. Biol.*, 8(5):483–491, 2001.

[5] A. Bergeron, P. Medvedev, and J. Stoye. Rearrangement models and single-cut operations. *J. Comput. Biol.*, 17(9):1213–1225, 2010.

[6] A. Bergeron, J. Mixtacki, and J. Stoye. A unifying view of genome rearrangements. In *Proc. 6th Workshop Algs. in Bioinf. (WABI'06)*, volume 4175 of *Lecture Notes in Comp. Sci.*, pages 163–173. Springer Verlag, Berlin, 2006.

[7] M. Blanchette, G. Bourque, and D. Sankoff. Breakpoint phylogenies. In S. Miyano and T. Takagi, editors, *Genome Informatics*, pages 25–34. Univ. Academy Press, Tokyo, 1997.

[8] G. Bourque and P.A. Pevzner. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.*, 12:26–36, 2002.

[9] L. Bulteau, G. Fertin, and I. Rusu. Sorting by transpositions is difficult. In *Proc. 38th Int'l Colloq. on Automata, Languages, and Programming (ICALP 2011)*, volume 6756 of *Lecture Notes in Comp. Sci.* Springer Verlag, Berlin, 2011.

[10] F. Burki, K. Shalchian-Tabrizi, M. Minge, A. Skjveland, S.I. Nikolaev, K.S. Jakobsen, and J. Pawlowski. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS ONE*, 2(8):e790, 2007.

[11] G. Cannarozzi, A. Schneider, and G. Gonnet. A phylogenomic study of human, dog, and mouse. *PLoS Comput. Biol.*, 3:e2, 2007.

[12] A. Caprara. Formulations and hardness of multiple sorting by reversals. In *Proc. 3rd Int'l Conf. Comput. Mol. Biol. (RECOMB'99)*, pages 84–93. ACM Press, New York, 1999.

[13] M.J. Chaisson, B.J. Raphael, and P.A. Pevzner. Microinversions in mammalian evolution. *Proc. Nat'l Acad. Sci., USA*, 103(52):19,824–19,829, 2006.

[14] F. Chen, A.J. Mackey, J.K. Vermunt, and D.S. Roos. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, 2(4):e383, 2007.

[15] W.H.E. Day. Computationally difficult parsimony problems in phylogenetic systematics. *J. Theoretical Biology*, 103:429–438, 1983.

[16] R. Desper and O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.*, 9(5):687–705, 2002.

[17] R. Desper and O. Gascuel. Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Mol. Biol. Evol.*, 21(3):587–598, 2003.

[18] R. Durrett, R. Nielsen, and T.L. York. Bayesian estimation of the genomic distance. *Genetics*, 166(1):621–629, 2004.

[19] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7(1):1–26, 1979.

[20] B. Efron. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589, 1981.

[21] B. Efron. The jackknife, the bootstrap and other resampling plans. In *CBMS-NSF Regional Conf. Series in Applied Math.*, volume 38. SIAM, 1982.

[22] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, 37(1):36–48, 1983.

[23] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

[24] N. El-Mabrouk. Genome rearrangement by reversals and insertions/deletions of contiguous segments. In *Proc. 11th Ann. Symp. Combin. Pattern Matching (CPM'00)*, volume 1848 of *Lecture Notes in Comp. Sci.*, pages 222–234. Springer Verlag, Berlin, 2000.

[25] M. A. Marra et al. The genome sequence of the sars-associated coronavirus. *Science*, 300(5624):1399–1404, 2003.

[26] U. Hellsten et al. The genome of the western clawed frog xenopus tropicalis. *Science*, 328(5978):633–636, 2010.

[27] J.S. Farris. The future of phylogeny reconstruction. *Zoologica Scripta*, 26(4):303–311, 1997.

[28] J.S. Farris, V.A. Albert, M. Källersjö, D. Lipscomb, and A.G. Kluge. Parsimony jackknifing outperforms neighbor-joining. *Cladistics*, 12(2):99–124, 1996.

[29] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[30] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evol.*, 39:783–791, 1985.

[31] J. Felsenstein. *Phylogenetic Inference Package (PHYLIP), Version 3.5*. University of Washington, Seattle, 1993.

[32] J. Felsenstein and H. Kishino. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Syst. Biol.*, 42(2):193–200, 1993.

[33] N.M. Ferguson, A.P. Galvani, and R.M. Bush. Ecological and immunological determinants of influenza evolution. *Nature*, 422:428–433, 2003.

[34] G. Fertin, A. Labarre, I. Rusu, E. Tannier, and S. Vialette. *Combinatorics of Genome Rearrangements*. MIT Press, 2009.

[35] D. Fitzpatrick, M. Logue, J. Stajich, and G. Butler. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology*, 6(1):99, 2006.

[36] O. Gascuel. *Mathematics of Evolution and Phylogeny*. Oxford Univ. Press, UK, 2005.

[37] P. Goloboff. Analyzing large datasets in reasonable times: solutions for composite optima. *Cladistics*, 15:415–428, 1999.

[38] S. Guindon and O. Gascuel. PHYML—a simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52(5):696–704, 2003.

[39] S. Hannenhalli and P.A. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proc. 27th Ann. ACM Symp. Theory of Comput. (STOC'95)*, pages 178–189. ACM Press, New York, 1995.

[40] S. Hannenhalli and P.A. Pevzner. Transforming mice into men (polynomial algorithm for genomic distance problems). In *Proc. 36th Ann. IEEE Symp. Foundations of Comput. Sci. (FOCS'95)*, pages 581–592. IEEE Press, Piscataway, NJ, 1995.

[41] M. Hasegawa, H. Kishino, and T.-A. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 21:160–174, 1985.

[42] D.M. Hillis and J.P. Huelsenbeck. Assessing molecular phylogenies. *Science*, 267:255–256, 1995.

[43] S. Holmes. Bootstrapping phylogenetic trees: theory and methods. *Statistical Science*, 18(2):241–255, 2003.

[44] F. Hu, N. Gao, M. Zhang, and J. Tang. Maximum likelihood phylogenetic reconstruction using gene order encodings. In *Proc. 2011 IEEE Symp. Comput. Intell. in Bioinf. & Comput. Biol. (CIBCB'11)*, pages 117–122. IEEE, 2011.

[45] J.P. Huelsenbeck and F. Ronquist. MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, 17:754b, 2001.

[46] D. Huson, S. Nettles, and T. Warnow. Disk-covering, a fast converging method for phylogenetic tree reconstruction. *J. Comput. Biol.*, 6(3):369–386, 1999.

[47] D. Huson and M. Steel. Phylogenetic trees based on gene content. *Bioinformatics*, 20(13):2044–2049, 2004.

[48] D. Huson, L. Vawter, and T. Warnow. Solving large scale phylogenetic problems using DCM-2. In *Proc. 7th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'99)*, 1999.

[49] G.A. Huttley, M.J. Wakefield, and S. Easteal. Rates of genome evolution and branching order from whole-genome analysis. *Mol. Biol. Evol.*, 24(8):1722–1730, 2007.

[50] H.T. Imai. On the origin of telocentric chromosomes in mammals. *J. Theor. Biol.*, 71(4):619–637, 1978.

[51] H.T. Imai and R.H. Crozier. Quantitative analysis of directionality in mammalian karyotype evolution. *American Naturalist*, 116(4):537–569, 1980.

[52] H.T. Imai, T. Maruyama, T. Gojobori, Y. Inoue, and R.H. Crozier. Theoretical bases for karyotype evolution. 1. the minimum-interaction hypothesis. *American Naturalist*, 128(6):900–920, 1986.

[53] H.T. Imai, Y. Satta, M. Wada, and N. Takahata. Estimation of the highest chromosome number of eukaryotes based on the minimum interaction theory. *J. Theor. Biol.*, 217(1):61–74, 2002.

[54] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[55] M. Kimura. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120, 1980.

[56] S. Kumar, K. Tamura, I.B. Jakobsen, and M. Nei. MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics*, 17(12):1244–1245, 2001.

[57] B. Larget, J.B. Kadane, and D.L. Simon. A Markov chain Monte Carlo approach to reconstructing ancestral genome arrangements. *Mol. Biol. Evol.*, 22:486–495, 2005.

[58] J.-F. Lefebvre, N. El-Mabrouk, E.R.M. Tillier, and D. Sankoff. Detection and validation of single gene inversions. In *Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03)*, volume 19 of *Bioinformatics*, pages i190–i196. Oxford U. Press, 2003.

[59] I. Letunic and P. Bork. Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucl. Acids Res.*, 39(suppl 2):W475–W478, 2011.

[60] Y. Lin and B.M.E. Moret. Estimating true evolutionary distances under the DCJ model. In *Proc. 16th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'08)*, volume 24(13) of *Bioinformatics*, pages i114–i122, 2008.

[61] Y. Lin and B.M.E. Moret. A new genomic evolutionary model for rearrangements, duplications, and losses that applies across eukaryotes and prokaryotes. *J. Comput. Biol.*, 18(9):1055–1064, 2011.

[62] Y. Lin, V. Rajan, and B.M.E. Moret. Bootstrapping phylogenies inferred from rearrangement data. In *Proc. 11th Workshop Algs. in Bioinf. (WABI'11)*, volume 6833 of *Lecture Notes in Comp. Sci.*, pages 175–187. Springer Verlag, Berlin, 2011.

[63] Y. Lin, V. Rajan, and B.M.E. Moret. Fast and accurate phylogenetic reconstruction from high-resolution whole-genome data and a novel robustness estimator. *J. Comput. Biol.*, 18(9):1130–1139, 2011.

[64] Y. Lin, V. Rajan, K.M. Swenson, and B.M.E. Moret. Estimating true evolutionary distances under rearrangements, duplications, and losses. In *Proc. 8th Asia Pacific Bioinf. Conf. (APBC'10)*, volume 11 (Suppl. 1):S54 of *BMC Bioinformatics*, 2010.

[65] M.D. López and T. Samuelsson. eGOB: Eukaryotic Gene Order Browser. *Bioinformatics*, 2011.

[66] M. Lynch. *The Origins of Genome Architecture*. Sinauer, 2007.

[67] J. Ma, A. Ratan, B.J. Raney, B.B. Suh, W. Miller, and D. Haussler. The infinite sites model of genome evolution. *Proc. Nat'l Acad. Sci., USA*, 105(38):14254–14261, 2008.

[68] G.M. Mace, J. L. Gittleman, and A. Purvis. Preserving the tree of life. *Science*, 300(5624):1707–1709, 2003.

[69] O. Madsen, M. Scally, C.J. Douady, D.J. Kao, R.W. DeBry, R. Adkins, H.M. Amrine, M.J. Stanhope, W.W. de Jong, and M.S. Springer. Parallel adaptive radiations in two major clades of placental mammals. *Nature*, 409:610–614, 2001.

[70] M. Marron, K.M. Swenson, and B.M.E. Moret. Genomic distances under deletions and insertions. *Theor. Computer Science*, 325(3):347–360, 2004.

[71] R.G. Miller. The jackknife-a review. *Biometrika*, 61(1):1, 1974.

[72] C. Mora, D.P. Tittensor, S. Adl, A.G.B. Simpson, and B. Worm. How many species are there on earth and in the ocean? *PLoS Biol*, 9(8):e1001127, 08 2011.

[73] B.M.E. Moret, J. Tang, L.-S. Wang, and T. Warnow. Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comput. Syst. Sci.*, 65(3):508–525, 2002.

[74] B.M.E. Moret, J. Tang, and T. Warnow. Reconstructing phylogenies from gene-content and gene-order data. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*, pages 321–352. Oxford Univ. Press, UK, 2005.

[75] B.M.E. Moret, L.-S. Wang, T. Warnow, and S.K. Wyman. New approaches for reconstructing phylogenies from gene-order data. In *Proc. 9th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'01)*, volume 17 of *Bioinformatics*, pages S165–S173, 2001.

[76] B.M.E. Moret and T. Warnow. Advances in phylogeny reconstruction from gene order and content data. In E.A. Zimmer and E.H. Roalson, editors, *Molecular Evolution: Producing the Biochemical Data, Part B*, volume 395 of *Methods in Enzymology*, pages 673–700. Elsevier, 2005.

[77] B.M.E. Moret, S.K. Wyman, D.A. Bader, T. Warnow, and M. Yan. A new implementation and detailed study of breakpoint analysis. In *Proc. 6th Pacific Symp. on Biocomputing (PSB'01)*, pages 583–594. World Scientific Pub., 2001.

[78] M.E. Mort, P.S. Soltis, D.E. Soltis, and M.L. Mabry. Comparison of three methods for estimating internal support on phylogenetic trees. *Syst. Biol.*, 49(1):160–171, 2000.

[79] W.J. Murphy, E. Eizirik, W.E. Johnson, Y.P. Zhang, O.A. Ryder, and S.J. O'Brien. Molecular phylogenetics and the origins of placental mammals. *Nature*, 409:614–618, 2001.

[80] E. Negrisolo, H. Kuhl, C. Forcato, N. Vitulo, R. Reinhardt, T. Patarnello, and L. Bargelloni. Different phylogenomic approaches to resolve the evolutionary relationships among model fish species. *Mol. Biol. Evol.*, 27(12):2757–2774, 2010.

[81] H. Nozaki, M. Matsuzaki, M. Takahara, O. Misumi, H. Kuroiwa, M. Hasegawa, T. Shin-i, Y. Kohara, N. Ogasawara, and T. Kuroiwa. The phylogenetic position of red algae revealed by multiple nuclear genes from mitochondria-containing eukaryotes and an alternative hypothesis on the origin of plastids. *J. Mol. Evol.*, 56:485–497.

[82] Aïda Ouangraoua, Frédéric Boyer, Andrew McPherson, Eric Tannier, and Cedric Chauve. Prediction of contiguous regions in the amniote ancestral genome. volume 5542 of *Lecture Notes in Comp. Sci.*, pages 173–185. Springer Verlag, Berlin, 2009.

[83] I. Pe'er and R. Shamir. The median problems for breakpoints are NP-complete. *Elec. Colloq. on Comput. Complexity*, 71, 1998.

[84] C.P. Ponting. The functional repertoires of metazoan genomes. *Nat. Rev. Genet.*, 9(9):689–698, 2008.

[85] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[86] D.R. Robinson and L.R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[87] A. Rokas and P.W.H. Holland. Rare genomic changes as a tool for phylogenetics. *Trends in Ecol. and Evol.*, 15:454–459, 2000.

[88] U. Roshan, B.M.E. Moret, T.L. Williams, and T. Warnow. Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. In *Proc. 3rd IEEE Comp. Systems Bioinf. Conf. CSB'04*, pages 98–109. IEEE Press, Piscataway, NJ, 2004.

[89] H.A. Ross and A.G. Rodrigo. Immune-mediated positive selection drives human immunodeficiency virus type 1 molecular variation and predicts disease duration. *Journal of Virology*, 76(22):11715–11720, 2002.

[90] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

[91] N. Salamin, M.W. Chase, T.R. Hodkinson, and V. Savolainen. Assessing internal support with large phylogenetic DNA matrices. *Mol. Phyl. Evol.*, 27(3):528, 2003.

[92] D. Sankoff and M. Blanchette. Multiple genome rearrangement and breakpoint phylogeny. *J. Comput. Biol.*, 5:555–570, 1998.

[93] D. Sankoff and M. Blanchette. Probability models for genome rearrangement and linear invariants for phylogenetic inference. In *Proc. 3rd Int'l Conf. Comput. Mol. Biol. (RECOMB'99)*, pages 302–309. ACM Press, New York, 1999.

[94] J. Shao and C.F.J. Wu. A general theory for jackknife variance estimation. *Annals of Statistics*, 17(3):1176–1197, 1989.

[95] J. Shi, Y. Zhang, H. Luo, and J. Tang. Using jackknife to assess the quality of gene order phylogenies. *BMC Bioinformatics*, 11(1):168, 2010.

[96] A.U. Sinha and J. Meller. Sensitivity analysis for reversal distance and breakpoint reuse in genome rearrangements. pages 37–48. World Scientific, 2008.

[97] B. Snel, P. Bork, and M.A. Huynen. Genome phylogeny based on gene content. *Nature Genetics*, 21(1):108–110, 1999.

[98] P.S. Soltis and D.E. Soltis. Applying the bootstrap in phylogeny reconstruction. *Statist. Sci.*, 18(2):256–267, 2003.

[99] M. Srivastava, E. Begovic, J. Chapman, N.H. Putnam, U. Hellsten, T. Kawashima, A. Kuo, T. Mitros, A. Salamov, M.L. Carpenter, A.Y. Signorovitch, M.A. Moreno, K. Kamm, J. Grimwood, J. Schmutz, H. Shapiro, I.V. Grigoriev, L.W. Buss, B. Schierwater, S.L. Dellaporta, and D.S. Rokhsar. The functional repertoires of metazoan genomes. *Nature*, 454(7207):955–960, 2008.

[100] A. Stamatakis. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[101] A. Stamatakis. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):2688–2690, 2006.

[102] M.A. Steel. The maximum likelihood point for a phylogenetic tree is not unique. *Syst. Biol.*, 43(4):560–564, 1994.

[103] A.H. Sturtevant. A crossover reducer in Drosophila melanogaster due to inversion of a section of the third chromosome. *Biol. Zent. Bl.*, 46:697–702, 1926.

[104] A.H. Sturtevant and G.W. Beadle. The relation of inversions in the x-chromosome of drosophila melanogaster to crossing over and disjunction. *Genetics*, 21:554–604, 1936.

[105] K.M. Swenson, M. Marron, J.V. Earnest-DeYoung, and B.M.E. Moret. Approximating the true evolutionary distance between two genomes. In *Proc. 7th SIAM Workshop on Algorithm Engineering & Experiments (ALENEX'05)*. SIAM Press, Philadelphia, 2005.

[106] D.L. Swofford. *PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b8*, 2001.

[107] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, B.K. Mable, and C. Moritz, editors, *Molecular Systematics*, pages 407–514. Sinauer Assoc., Sunderland, MA, 1996.

[108] D.L. Swofford, G.J. Olsen, P.J. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, B.K. Mable, and C. Moritz, editors, *Molecular Systematics*, pages 407–514. Sinauer Assoc., Sunderland, MA, 1996.

[109] D.L. Swofford, G. Olson, P. Waddell, and D.M. Hillis. Phylogenetic inference. In D.M. Hillis, C. Moritz, and B. Mable, editors, *Molecular Systematics, 2nd ed.*, chapter 11. Sinauer Assoc., Sunderland, MA, 1996.

[110] J. Tang and B.M.E. Moret. Scaling up accurate phylogenetic reconstruction from gene-order data. In *Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03)*, volume 19 of *Bioinformatics*, pages i305–i312. Oxford U. Press, 2003.

[111] E. Tannier, C. Zheng, and D. Sankoff. Multichromosomal genome median and halving problems. In *Proc. 8th Workshop Algs. in Bioinf. (WABI'08)*, volume 5251 of *Lecture Notes in Comp. Sci.*, pages 1–13. Springer Verlag, Berlin, 2008.

[112] W. Thuiller, S. Lavergne, C. Roquet, I. Boulangeat, B. Lafourcade, and M. B. Araujo. Consequences of climate change on the tree of life in europe. *Nature*, 470:531–534, 2011.

[113] J.-N. Volff and J. Altenbuchner. A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol. Lett.*, 186:143–150, 2000.

[114] H. Wang, Z. Xu, L. Gao, and B. Hao. A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC Evolutionary Biology*, 9(1):195, 2009.

[115] L.-S. Wang. Exact-IEBP: a new technique for estimating evolutionary distances between whole genomes. In *Proc. 33rd Ann. ACM Symp. Theory of Comput. (STOC'01)*, pages 637–646. ACM Press, New York, 2001.

[116] L.-S. Wang, R.K. Jansen, B.M.E. Moret, L.A. Raubeson, and T. Warnow. Fast phylogenetic methods for genome rearrangement evolution: An empirical study. In *Proc. 7th Pacific Symp. on Biocomputing (PSB'02)*, pages 524–535. World Scientific Pub., 2002.

[117] L.-S. Wang and T. Warnow. Estimating true evolutionary distances between genomes. In *Proc. 1st Workshop Algs. in Bioinf. (WABI'01)*, number 2149 in Lecture Notes in Comp. Sci., pages 176–190. Springer Verlag, Berlin, 2001.

[118] D.E. Wildman, M. Uddin, J.C. Opazo, G. Liu, V. Lefort, S. Guindon, O. Gascuel, L.I. Grossman, R. Romero, and M. Goodman. Genomics, biogeography, and the diversification of placental mammals. *Proc. Nat'l Acad. Sci., USA*, 104(36):14395–14400, 2007.

[119] S. Yancopoulos, O. Attie, and R. Friedberg. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21(16):3340–3346, 2005.

[120] S. Yancopoulos and R. Friedberg. Sorting genomes with insertions, deletions and duplications by DCJ. In *Proc. 6th RECOMB Workshop Comp. Genomics (RECOMB-CG'08)*, volume 5267 of *Lecture Notes in Comp. Sci.*, pages 170–183. Springer Verlag, Berlin, 2008.

[121] T.L. York, R. Durrett, and R. Nielsen. Dependence of paracentric inversion rate on tract length. *BMC Bioinformatics*, 8(115), 2007.

[122] H. Zhang, Y. Zhong, B. Hao, and X. Gu. A simple method for phylogenomic inference using the information of gene content of genomes. *Gene*, 441:163–168, 2009.

[123] E. Zuckerkandl and L.B. Pauling. *Molecular disease, evolution, and genetic heterogeneity*, pages 189–225. Academic Press, New York, 1962.

CURRICULUM VITAE

Yu Lin

Contact Information

EPFL IC IIF LCBB
INJ 211 Station 14
CH-1015, Lausanne
Switzerland

Phone: (+41)76 65 02468
Fax: (+41)21 69 37555
E-mail: yu.lin@epfl.ch

Research Interests

Computational Biology and Bioinformatics:
Comparative genomics: models and distance estimations in whole-genome evolution.
Computational phylogenetics: algorithms and statistical inference in phylogenetic reconstruction based on whole-genome data.
Computational proteomics: models and algorithms of peptide identifications through mass spectrometry.

Education

Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
Ph.D., Computer Science, Aug. 2007 – Jul. 2012
- Advisor: Prof. Bernard Moret

Chinese Academy of Sciences(CAS), Beijing, China
M.S., Computer Science, Sep. 2004 – Jul. 2007
- Advisor: Prof. Dongbo Bu

University of Science and Technology of China (USTC), Hefei, China
B.S., Computer Science, Sep. 2000 – Jul. 2004

Research Experience

Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Research Assistant in comparative genomics and computational phylogenetics
- Advisor: Prof. Bernard Moret Aug. 2007 – Jul. 2012

Chinese Academy of Sciences, China
Research Assistant in computational proteomics
- Advisor: Prof. Dongbo Bu Sep. 2006 – Jul. 2007

City University of Hong Kong, Hong Kong, China
Research Assistant in biclustering algorithms for biological data
- Advisor: Prof. Lusheng Wang Dec. 2005 – Aug. 2006

Teaching Experience

Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland,

Teaching Assistant: Advanced Algorithms (Fall 2009, Fall 2010), Topics in Bioinformatics I (Fall 2007, Fall 2008), Computational Molecular Biology (Spring 2008).

Guest Lecturer: Advanced Algorithms (Fall 2011), Advanced Theoretical Computer Science (Spring 2009), Computational Molecular Biology (Spring 2009).

Professional Activities:

Referee
Journal of Combinatorial Optimization, PLoS One,
RECOMB'10, WABI'11, WABI'12, ALENEX'12

Honors

Chinese Government Award For Outstanding Self-Financed Students Abroad, 2012
Director's Prize, ICT, Chinese Academy of Sciences, 2007
Guo Moruo Prize, University of Science and Technology of China, 2003

Publications

Shao, M., Lin, Y., "Approximating the edit distance for genomes under DCJ, insertion and deletion," *Proc. 10th RECOMB Workshop on Comparative Genomics (RECOMB-CG'12)*, accepted, to appear in *BMC Bioinformatics*.

Lin, Y., Rajan, V., and Moret, B.M.E., "Bootstrapping phylogenies inferred from rearrangement data," *BMC Algorithms for Molecular Biology*, accepted, to appear.

Lin, Y., Rajan, V., and Moret, B.M.E., "A metric for phylogenetic trees based on matching," *IEEE/ACM Trans. on Computational Biology and Bioinformatics* 9, 4, 1014-1022 (2012).

Lin, Y., Rajan, V., and Moret, B.M.E., "Bootstrapping phylogenies inferred from rearrangement data," *Proc. 11th Workshop on Algorithms in Bioinformatics (WABI'11)*, Lecture Notes in Computer Science 6833, 175-187, Springer Verlag (2011).

Lin, Y., and Moret, B.M.E., "A new genomic evolutionary model for rearrangements, duplications, and losses that applies across eukaryotes and prokaryotes," *J. Computational Biology* 18, 9, 1055-1064 (2011).

Lin, Y., Rajan, V., and Moret, B.M.E., "Fast and accurate phylogenetic reconstruction from high-resolution whole-genome data and a novel robustness estimator," *J. Computational Biology* 18, 9, 1131-1139 (2011).

Lin, Y., Rajan, V., and Moret, B.M.E., "A metric for phylogenetic trees based on matching," *Proc. 7th Int'l Symp. Bioinformatics Research & Appls. (ISBRA'11)*, Lecture Notes in Computer Science 6674, 197-208, Springer Verlag (2011).

Lin, Y., and Moret, B.M.E., "A new genomic evolutionary model for rearrangements, duplications, and losses that applies across eukaryotes and prokaryotes," *Proc. 8th RECOMB Workshop on Comparative Genomics (RECOMB-CG'10)*, in Lecture Notes in Computer Science 6398, 228-239, Springer Verlag (2010).

Lin, Y., Rajan, V., and Moret, B.M.E., "Fast and accurate phylogenetic reconstruction from high-resolution whole-genome data and a novel robustness estimator," *Proc. 8th RECOMB Workshop on Comparative Genomics (RECOMB-CG'10)*, in Lecture Notes in Computer Science 6398, 137-148, Springer Verlag (2010).

Lin, Y., Rajan, V., Swenson, K.M., and Moret, B.M.E., "Estimating true evolutionary distances under rearrangements, duplications, and losses," *Proc. 8th Asia-Pacific Bioinformatics Conf. (APBC'10)*, in *BMC Bioinformatics 2010*, 11 (Suppl. 1):S54.

Rajan, V., Xu, A.W., Lin, Y., Swenson, K.M., and Moret, B.M.E., "Heuristics for the inversion median problem," *Proc. 8th Asia-Pacific Bioinformatics Conf. (APBC'10)*, in *BMC Bioinformatics* (Suppl. 1):S30 (2010).

Swenson, K.M., Rajan, V., Lin, Y., and Moret, B.M.E., "Sorting signed permutations by inversions in O(nlogn) time," *J. Computational Biology*, 17, 3 (2010), 489-501.

Swenson, K.M., Lin, Y., Rajan, V., and Moret, B.M.E., "Hurdles and sorting by inversions: Combinatorial, statistical, and experimental results," *J. Computational Biology*, 16(10):1339-1351 (2009).

Swenson, K.M., Rajan, V., Lin, Y., and Moret, B.M.E., "Sorting signed permutations by inversions in O(nlogn) time," *Proc. 13th Int'l Conf. on Research in Comput. Molecular Biol. (RECOMB'09)*, in Lecture Notes in Computer Science 5541, 386-399, Springer Verlag (2009).

Swenson, K.M., Lin, Y., Rajan, V., and Moret, B.M.E., "Hurdles hardly have to be heeded," *Proc. 6th RECOMB Workshop on Comparative Genomics (RECOMB-CG'08)*, in Lecture Notes in Computer Science 5267, 239-249, Springer Verlag (2008).

Lin, Y., and Moret, B.M.E., "Estimating true evolutionary distances under the DCJ model", *Proc. 16th Conf. on Intelligent Systems for Molecular Biol. (ISMB'08)*, in *Bioinformatics* 24(13):i114-i122 (2008).

Wang, L., Lin, Y., and Liu, X., "Approximation algorithms for biclustering problems", *SIAM J. Computing* 38(4): 1504-1518 (2008).

Lin, Y., Qiao, Y., Sun, S., Yu, C, Dong G., and Bu, D. "A fragmentation event model for peptide identication by mass spectrometry," *Proc. 12th Int'l Conf. on Research in Comput. Molecular Biol. (RECOMB'08)*, in Lecture Notes in Computer Science 4955, 154-166, Springer Verlag (2008).

Sun, S., Yu, C., Qiao, Y., Lin, Y., Dong, G., Liu, C., Zhang, J., Zhang, Z., Cai, J., Zhang, H., and Bu, D., "Deriving the probabilities of water loss and ammonia loss for amino acids from tandem mass spectra", *J. Proteome Research,* 7 (01): 202-208 (2008).

Yu, C., Lin, Y., Sun, S., Cai, J., Zhang, J., Bu, D., Zhang, Z., and Chen, R., "An iterative algorithm to quantify factors influencing peptide fragmentation during tandem mass spectrometry", *J. Bioinformatics and Computational Biology,* 5(2a):297-311 (2007).

Wang, L., Lin, Y., and Liu, X., "Approximation algorithms for bi-clustering problems," *Proc. 6th Workshop on Algorithms in Bioinformatics (WABI'06)*, Lecture Notes in Computer Science 4175, 310-320, Springer Verlag (2006).

Yu, C., Lin, Y., Sun, S., Cai, J., Zhang, J., Bu, D., Zhang, Z., and Chen, R., "An iterative algorithm to quantify the factors influencing peptide fragmentation for MS/MS spectrum", *Proc. 5th Comput. Systems Bioinformatics Conf. (CSB'06)*, 353-360, Imperial College Press (2006)