Inference of Insertion and Deletion Scenarios for Ancestral Genome Reconstruction and Phylogenetic Analyses: Algorithms and Biological Applications

Abdoulaye Baniré Diallo

Doctor of Philosophy

School of Computer Science

McGill University

Montreal, Quebec, Canada

January 2009

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

This thesis is dedicated to my late grand fathers who inspired the infinite quest for knowledge in my life.

ACKNOWLEDGEMENTS

Since I joined the McGill Centre for Bioinformatics under Prof. Mathieu Blanchette's supervision, I have worked with the best people that I would have never expect to work with. Those are people whose valuable contributions in several ways has permit the completion of this thesis. I will never forget the long and productive lab meetings and discussions. I would first like to thank my advisor, Prof. Mathieu Blanchette for giving me the great opportunity to work in his group and to present my work at several scientific meetings. I am very grateful for the guidance, the encouragement, the trust, and the intellectual freedom he provided throughout my three-year tenure in his laboratory. I was very the lucky to have such a wiseman as supervisor.

Second, I will never forget all the help, advises, discussions and availability of my co-supervisor Vladimir Makarenkov. He acts as a real mentor for me. He interested me in doing my master and Ph.D. Here, I would like to present all my gratitude for all the things that he has done for me the last five years. I am also grateful to the members of my Ph.D. committee, Prof. Mike Hallet and Prof. Ted Perkins who have contributed with their valuable comments and ideas from the very beginning of my Ph.D. research.

Thanks to my collaborators and friends Alpha Boubacar Diallo, Alix Boc, Dunarel Badescu, Emmanuel Mongin, Éric Gaul, Michael Mayhew, Van Dung Nguyen, Wessam Ajib for their friendship, unselfish support, assistance, and interest in my projects. All our discussions in the soccer fields, the labs and the clubs have made this thesis reality. All my gratitudes to the Potvin family for their unconditional support and acceptance as one of them. My best regards to my parents (Dr. Aliou Baniré Diallo and Dr. Hawa Thiam) for their unconditional support and unambiguous love throughout my life. Their love

for knowledge and science give me all the courage in my studies.

I give my heartfelt thanks to my wife Ramatoulaye Bah for her patience and help, to our wonderful son Aliou Baniré Diallo for giving a reason to my life and my brother Sekou Hamid Baniré Diallo for being always there for me.

Finally, I am thankful to the Natural Sciences and Engineering Research Council of Canada (NSERC) for funding my Ph.D. project.

For those I omit to mention, find here my sincerest gratitude for all the help, and attention you provide to me. My late grand father used to say: "to be a good man, have the inspiration from all people around you and more from people that are far from you". Hence, Thanks to all people that inspired me and continue to inspire me.

ABSTRACT

This thesis focuses on algorithms related to ancestral genome reconstruction and phylogenetics analyses. Specially, it studies insertion and deletion (indel) in genomic sequences, their utilities for (1) evolutionary studies of species families, (2) multiple alignment and phylogenetic trees reconstruction assessment, and (3) functional DNA sequence annotation. Here, the indel scenarios reconstruction problem is presented, in a likelihood framework, and it can be stated as follows: given a multiple alignment of orthologous sequences and a phylogenetic tree for these sequences, reconstruct the most likely scenario of insertions and deletions capable of explaining the gaps observed in the alignment. This problem, that we called the Indel Maximum Likelihood Problem (IMLP), is an important step toward the reconstruction of ancestral genomic sequences, and is important for studying evolutionary processes, genome function, adaptation and convergence.

In this thesis, first, we showed that we can solve the IMLP using a new type of tree hidden Markov model whose states correspond to single-base evolutionary scenarios and where transitions model dependencies between neighboring columns. The standard Viterbi and Forward-backward algorithms are optimized to produce the most likely ancestral reconstruction and to compute the level of confidence associated to specific regions of the reconstruction. A heuristic is presented to make the method practical for large data sets, while retaining an extremely high degree of accuracy. The developed methods have been made available for the community through a web interface. Second we showed the utilities of the defined indel score for assessing the accuracy of multiple sequence alignment and phylogenetic tree reconstruction. Third, the

provided method is included into the framework of the ancestral protein reconstruction of phages under a reticulate evolution and the evolutionary studies of the carcinogencity of the Human Papilloma Virus family.

The results presented in this thesis contribute in different areas of research such as multiple sequence alignment refinement, agreement between phylogenetic trees and related multiple sequences alignment, analysis of evolutionary processes and many other problems related to comparative genomics.

ABRÉGÉ

Cette thèse traite d'algorithmes pour la reconstruction de génomes ancestraux et l'analyse phylogénétique. Elle étudie particulièrement les scénarios d'insertion et délétion (indels) dans les séquences génomiques, leur utilité (1) pour l'étude des familles d'espèces, (2) pour l'évaluation des alignements multiples de séquences et la reconstruction phylogénétique, (3) et pour l'annotation de séquences génomiques fonctionnelles. Dans cette thèse, le problème de la reconstruction du scénario d'indels est étudié en utilisant le critère de maximum de vraisemblance. Ce problème peut être défini de la manière suivante: étant donné un alignement multiple de séquences orthologues et un arbre phylogénétique traduisant l'histoire évolutive de ces séquences, reconstruire le scénario d'indels le plus vraisemblable capable d'expliquer les brèches présentes dans l'alignement. Ce problème, dénommé "Indel Maximum Likelihood Problem (IMLP)", est une importante étape de la reconstruction de séquences ancestrales. Il est également important pour l'étude des processus évolutifs, des fonctions des gènes, de l'adaptation et de la convergence.

Dans une première étape de cette thèse, nous montrons que l'IMLP peut être résolu en utilisant un nouveau type de données combinant un arbre phylogénétique et un modèle de Markov caché. Les états de ce modèle de Markov caché correspondent à un scénario évolutif d'une colonne de l'alignement. Ses transitions modélisent la dépendance entre les colonnes voisines de l'alignement. Les algorithmes standard de Viterbi et de Forward-Backward ont été optimisés pour produire le scénario ancestral le plus vraisemblable et pour calculer le niveau de confiance associé aux prédictions. Dans cette thèse, Nous présentons également une heuristique qui permet d'adapter la méthode à des données de grandes tailles. En second, nous montrons l'utilité du score

d'indel dans l'évaluation d'alignement multiple de séquences et de reconstruction d'arbres phylogénétiques. Troisièmement, la méthode proposée a été implantée dans le projet de reconstruction de séquences protéiques ancestrales des bactériophages qui ont une évolution réticulée. Elle a également été utilisée dans l'étude de l'évolution de la carcinogénécité des virus du Papillome Humain.

Les résultats présentés dans cette thèse permettent d'ouvrir la voie sur plusieurs problèmes comme la correction des erreurs d'alignement, l'étude de la phylogénie conjointe à l'alignement multiple de séquences, l'analyse de processus évolutifs et bien d'autres problèmes de la génomique comparée.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Genome organization and evolution

One of the major outcomes of the Darwinian Theory [43] is that species evolve through changes occurring over time. The independent changes in the genomic patrimonies of living species lead to the rise of new organisms. Existing organisms range from bacteria to multicellular plants and animals. There are three major recognized domains of life: Eubacteria (e.g. Cyanobacteria, Spirochetes, etc.), Archea or Archeabacteria (e.g. Crenarchaeotes, etc.) and Eukaryotes (e.g. animals, plants, etc.). Species belonging to the first two domains are called prokaryotes. Their cells do not have a true nucleus and their DNA is not structured as eukaryotic chromosomes. For instance, prokaryotes have cell membranes and cytoplasm, but their DNA is not separated from the cytoplasm by a nuclear membrane as Eukaryotes. Eukaryotic species can be divided as having a single cell (unicellular) or multiple cells (multicellular). Most of the eukaryotes have mitochondria, where the major steps in aerobic respiration occur [49]. Moreover, plant cells may have chloroplasts for photosynthesis.

There are several similarities between organisms of different species, such as the presence of the cell as basic unit of life, the DNA molecules encoding the genetic information, the transcription of the information into RNA and its translation into protein, the gene distribution and organization, and the function of different genomic regions [49]. Their shared properties reflect the common origin of life depicted by their evolutionary relationships. However, the variation in the sequence of nucleotides contained on chromosomal (or

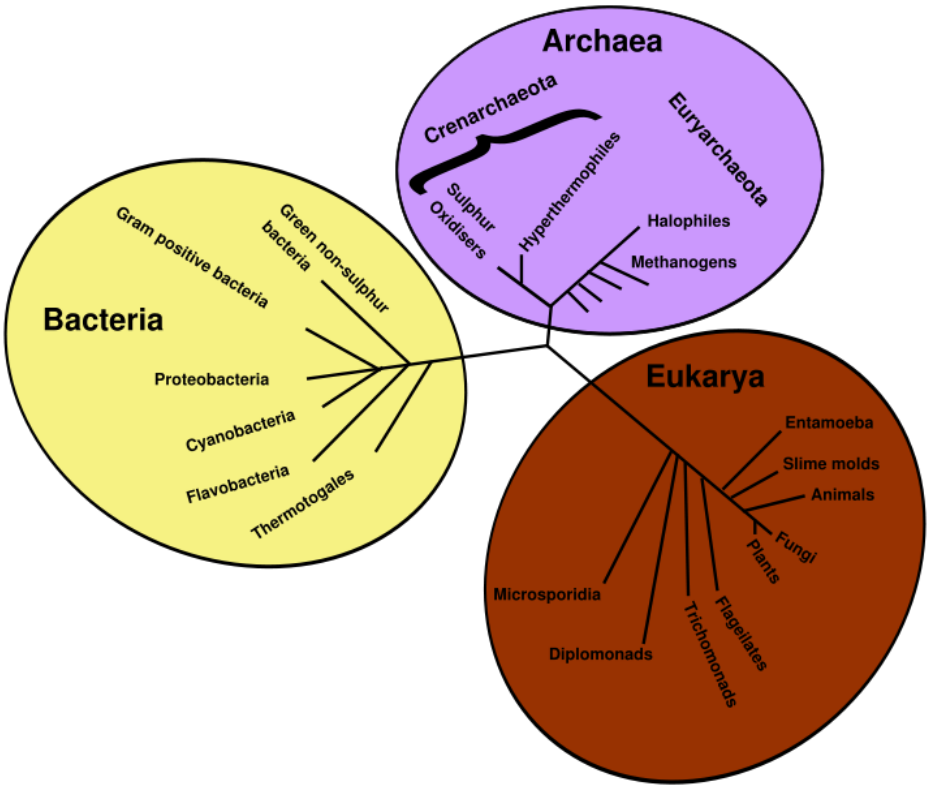


Figure 1–1: The living organisms are divided in three major domains. Each circle represents one of the domains with a sample of related species. This figure is taken from [238].

non-chromosomal) DNA molecules are the result of sets of evolutionary events that have taken place. These events, called ***mutations*** are linked to how DNA is copied (due to errors in the copying process or replication process) and how chromosomes are recombined. The consequences of these events are the diversity of the observed genomes. The goal of comparative genomics is to analyze the function of regions by comparing the genomic data of different individuals or species. The shared regions that derive from a common ancestor are called ***homologous.*** However, comparing the genomic sequences of living organisms leads to the observation of either homologous DNA sequences due to the evolutionary relationship or mutations that cause differences between the organisms. Hence, comparing organisms can help reveal the origin of phenotypic and functional divergence between species. The different types of

mutations can be divided in two parts (small-scale and large-scale mutations) presented in subsections 1.1.1 and 1.1.2.

1.1.1 Small scale mutations

The genome is made up of a concatenation of DNA bases that are composed of four different nucleotides: Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). Small scale mutations affect either a single nucleotide or a small region that does not cross chromosome boundaries. The following mutations affecting DNA are considered small scale mutations:

- **Substitutions or point mutations:** a single DNA base can flip into another base due to similarity of their chemical structure. Substitutions are the most common mutations that occur during genomes evolution. In fact, the adenine (A) molecule is similar in structure to the guanine (G) molecule (they are called purines), and the thymine (T) molecule is similar to the cytosine (C) one (they are called pyrimidines). A purine is more likely to change to another purine and a pyrimidine is more likely to change to a pyrimidine. Mutations between purines or pyrimidines are called transitions while the others are called transversions, as shown in Figure 1–2.

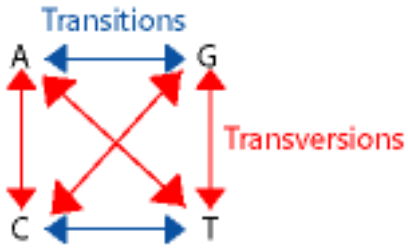


Figure 1–2: The types of point mutations. The blue lines (transitions) are the substitutions that are the most likely to occur.

- **Insertions and Deletions:** An insertion corresponds to the addition of one or more contiguous bases between adjacent nucleotides in a DNA sequence. Large insertions are often associated with transposable

elements [49]. The term deletion refers to the elimination of one or several contiguous nucleotides during the DNA copying (replication phase). However, when comparing the genomic data of two extant organisms, it is impossible to distinguish between insertions and deletions (see Figure 1–3). This ambiguity can only be solved by knowing the sequence of the common ancestor of the compared genomes. Thus, the terminology **indel** is used to denote either an insertion or a deletion. More details about indels are presented in Chapter 2.

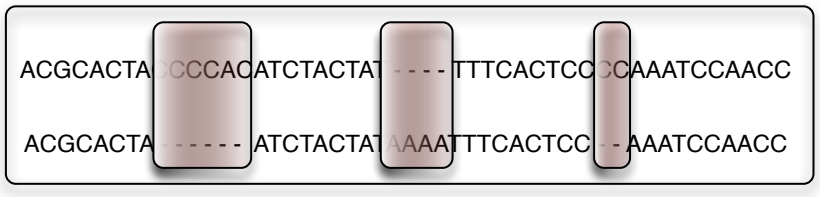


Figure 1–3: Indels representation in genome comparison. The symbol (-), called gap, is used to denote the absence of corresponding character at the given position of the comparison.

1.1.2 Large scale mutations

Large scale mutations are due to chromosomal reorganizations or to changes affecting large portions of DNA sequences. The following large scale mutations can be found when comparing genomes:

- **Duplication:** it results in several copies of the same portion of the genome being present either in different locations (**Segmental duplication**) or contiguously (**Tandem duplication**). The duplicated regions could be arge regions containing several genes. It could even involve the whole genome (whole genome duplication).
- **Inversion** (also called **Reversal**): it corresponds to the replacement of one strand of a DNA region with its complement while the 5' to 3' polarity is unchanged [49]. It reverses the order of the genes or markers present in the related DNA region (see Figure 1–4).

- **Transposition:** it consists of cutting out a genomic region and inserting it elsewhere in the genome (see Figure 1–4). The inserted segment could be also reversed prior to the insertion.
- **Translocation:** it is a cut in two different chromosomes and followed by a fusion of the obtain segments with the rest of the chromosomes such that each segment that has been cut will be reinserted in different chromosome (see Figure 1–4).
- **Chromosome fission and fusion:** it allows the genome to respectively increase or reduce the number of chromosomes present in its structure by breaking one chromosome into two pieces, or joining two chromosomes into one.

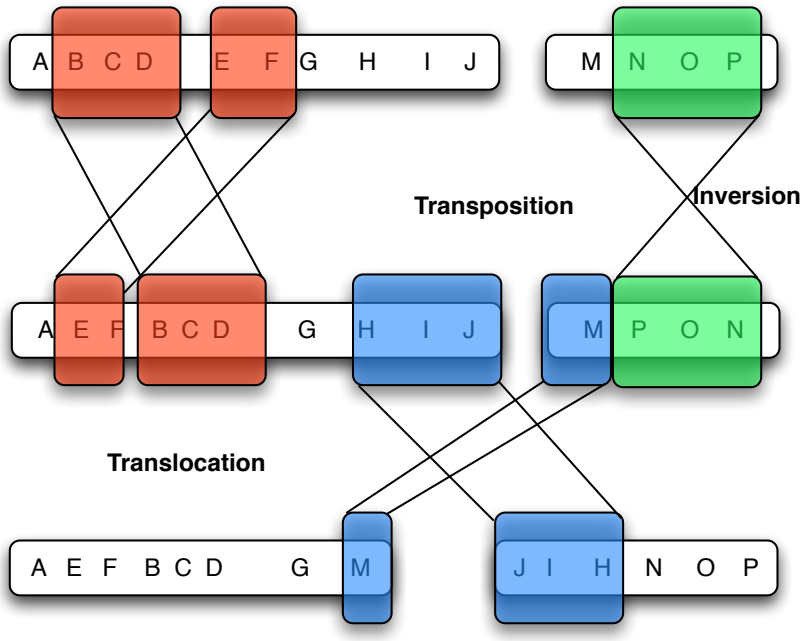


Figure 1–4: Inversion, transposition and translocation of genomic regions consisting of two chromosomes. Letters represent markers or genes, and boxes represent chromosomes.

Such mutations are called genome rearrangements and they are less frequent than small scale mutations. Recent progress in genome-scale sequencing and comparative mapping has increased the number of genome rearrangement

studies. The problem can be defined as finding the ancestral genome organization according to the genome organization of extant organisms [173]. This problem is hard to solve, however it has received a lot of attention (e.g. [20, 22, 231]).

1.2 Identification of homologous genomic regions

One important basis of comparative genomics is the identification of homologous regions based on sequence similarity. In fact, sequence similarity is often linked to function into the biological processes of organisms. To measure similarity between homologous sequences, they have to be aligned. Sequence alignment is one of the most studied fields in computational biology [1, 2, 112, 180, 209, 222]. Alignment can be either ***pairwise*** [1, 212], comprising just two sequences, or ***multiple*** [17, 25, 65, 112], with an arbitrary number of sequences. The main motivation for pairwise alignment relies on the inference of function while for multiple sequence alignment, it relies on the research of common sequence features between homologous sequences that share the same function. In fact, pairwise alignment could be considered as a special case of multiple sequence alignment, but in pratice the complexity of multiple sequence alignment is such that it could not be a straight extension of the pairwise method [106]. In most cases, multiple sequence alignments are built by repeatedly merging pairwise sequence alignments [25, 27, 112]. In 1.2.1 and 1.2.2, we will present the most common approaches for pairwise and multiple sequence alignment.

1.2.1 Pairwise alignment

The pairwise alignment problem can be defined as follows: Given two genomic sequences, find the *one-to-one ordered* correspondence between each

character in the two sequences such that a given criterion is optimized (minimizing substitutions, optimizing a distance or similarity score, the most likelihood one-to-one relationship, etc.). Pairwise alignment methods are widely used in comparative genomics, for example for retrieving data from databases such as ***Genbank*** [96], and identifying the function of genomic regions [1]. There are two main approaches to align genomic data:

- **Global Alignment:** this approach is preferred when the alignment is done for sequences that share homology for the entire sequence lengths. Whole sequences given as input have to be aligned in a single alignment where homologous regions have to be co-linear, with no genome rearrangement and no duplication. As mentioned in [14], global alignment is adequate when the genomic regions to be aligned derive from common ancestor with only small scale mutation events. The latter genomic regions are called ***orthologous*** [15]. Most of the global pairwise alignment methods derive from the widely used ***Needleman-Wunsch*** dynamic programming algorithm [180]. A list of popular global alignment methods includes: ***Dialign*** [170], ***MUMmer*** [47], ***Avid*** [24] and ***LAGAN*** [27].
- **Local Alignment:** this method tries to align a part of small different segments of the given genomic regions. This is motivated by the fact that many genomic sequences only share some homologous regions while other parts of the sequences are unrelated [49, 106]. Hence, a local alignment provides a prediction of the homology for a pair of subregions of the genomic regions to be aligned. Local alignments can help to predict homologous regions in genomic sequences that have undergone rearrangement, and it is the best choice when the sequences to be aligned are highly divergent (presence of a large fraction of unalignable regions)

[38, 14]. Most of the local pairwise alignment methods derive from the well-known ***Smith-Waterman*** dynamic programming algorithm [222]. There are a lot of local alignment methods. The most commonly used are ***BLAST*** [1] and ***BLASTZ*** [212].

Although there is apparently strict separation between the different types of pairwise alignment, most of the recent methods combine the global and local strategies to find accurate alignment. Those methods identify the sets of local alignments and then assemble them into a chain of co-linear blocks of local alignment [26, 14].

1.2.2 Multiple sequence alignment

Multiple sequence alignment is an extension of the pairwise alignment to more than two genomic sequences. These alignments are ubiquitous in molecular biology, particularly in comparative genomics [106]. Multiple sequence alignments contribute to genome annotation through techniques for finding homologies between sequence families [149, 172, 207], identifying and characterizing gene regions [51, 69, 137], etc. For these purposes, multiple sequence alignments are often modeled as profiles [73, 74] or as hidden Markov models [88, 136, 137]. Another important contribution of the multiple sequence alignment concerns the prediction of secondary and tertiary protein structure [9, 114, 77]. In proteins, residues have different evolution schemes depending on their role in the protein structure. Hence, the analysis of multiple sequence alignments by looking at the distribution of substitutions at each position procures an information about the protein structure [183]. Moreover, in RNA, the identification of correlated mutations is the basis of predicting secondary structure [114, 254]. Finally, multiple alignments are also the core of phylogenetic analyses [75, 85, 174, 229]. While phylogenetic distance methods allow one to compute evolutionary distances based on pairwise alignments, in

character based methods, the columns of multiple sequence alignments are individually correlated to a phylogenetic tree. More details about phylogenetic trees are presented in section 1.3.

The definition of a *good* alignment varies according to the type of studies. Hence, in the literature, good alignment satisfies one of the following criterion:

- ***Mathematical objective functions:*** These are commonly based on stochastic models of sequence evolution specifying the probability of events such as substitutions, insertions and deletions [14]. Using those models, alignment approaches try to maximize either the sum of the probabilities of all the evolutionary scenarios that can lead to the alignment, or a weighted/unweighted sum-of-pairs of aligned nucleotides [67].
- ***Phylogenetic correctness:*** Here we try to align together nucleotides that share the same ancestor. Hence, we try to reproduce the evolutionary history of each nucleotide.
- ***Function-based:*** This is often used when trying to align together regions that are related according to their functional similarity such as transcription factor binding site [14]. Here, aligned regions could be merely different in the nucleotide-level.

There exist several computational techniques for computing multiple sequence alignments. Each technique is adapted to how the obtained alignment will be used. Several methods extend the definition of a pairwise alignment to multiple sequence alignment. In this case, several sequences can be aligned together using an $n-$dimensional dynamic programming approach [170, 171]. However such a method is limited to small datasets. Instead of computing directly a multiple sequence alignment, several methods such as Clustal [112], Mavid [25], Mlagan [27], Muscle [71], and TBA [17] perform multiple pairwise alignments. The combination of the obtained pairwise alignment to a

multiple sequence alignment is performed in different ways according to the existing methods. Most of those alignment methods use a (user provided or precomputed) ***guide tree*** to determine the order in which the pairwise alignments will be merged [112, 25, 17]. This technique is called ***progressive alignment.*** One can notice that multiple methods are used to align the pairwise alignments such as computing ancestral sequences [25], computing profiles [73, 72, 112] or computing the sum-of-pairs [27, 102]. To increase the accuracy of progressive alignment heuristics, the alignment procedure can be iterated a number of times. The iteration procedure helps to correct mistakes introduced earlier in the pairwise alignment computation and to refine the whole alignment[246, 248]. Actually, iterated methods such as Muscle [71] and ProbCons [65] tend to give more accurate alignments.

Due to the importance of multiple sequence alignment, many other different computational approaches have been used. Genetic algorithms have been used to optimize multiple sequence alignment given an objective function [182, 183, 246]. Methods such as POA [143] build the multiple sequence alignments using partial order graphs and MSA [102] tries to find an optimal multiple sequence alignment using the branch and bound technique [246]. Multiple sequence alignment algorithms based on fast Fourier transform have also been built to improve the time complexity [129]. Finally, there is an increasing interest for probabilistic multiple sequence alignment. This new objective leads to the emergence of a bayesian approach to multiple sequence alignment [115] and statistical alignment [108].

1.3 Phylogenetic Trees and Networks

A phylogenetic tree is a classical way to illustrate species evolutionary relationships. The tree can be either rooted or not, as shown in Figure 1–5. The leaves of the phylogenetic tree correspond to the species that provided the

information for the tree inference. When the phylogenetic tree is rooted, the internal nodes represent hypothetical ancestor of the related subtree, while the overall root refers to the common ancestor of the species represented. Today phylogenetic trees are build mostly from molecular sequences (DNA, RNA and proteins). All existing approaches for building trees take in input multiple aligned sequences, with the exception of new emerging tools dedicated to small datasets and using unaligned sequences [72, 108, 109].

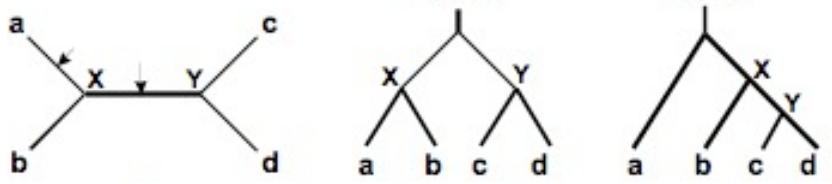


Figure 1–5: Unrooted phylogenetic tree (left tree) can be rooted in one of its branches, as shown in the two rightmost trees. Taken from Diallo et al. [52].

To reconstruct the evolutionary history of a given aligned sequences, two main approaches exist. For details on the methods for inferring phylogenetic trees, readers are referred to [75, 87, 228]. These two main approaches are:

- ***The distance based approach*** does not make reference to an historical relationship. It computes from aligned molecular sequences, pairwise distances between sequences. Then, a hierarchical clustering procedure allows to reconstruct the phylogenetic tree from distances. Those methods tend to be rapid but lack on consistency (due to the fact a large part of the information contained in the DNA sequence structure is lost when it is transformed into a distance).
- ***The character based approach*** is based on genealogy. It finds optimal trees by applying evolutionary models to constitute features of the ancestor at each node. The maximum parsimony [90], maximum likelihood [82] and Bayesian methods [196] belong to this approach.

The popular software packages PHYLIP [85] and PAUP [229] implement the distance, parsimony and likelihood methods of inference and also provide visualization tools and tree validation techniques. There exists several other phylogenetic tree reconstruction softwares including MEGA [138], DAMBE [255], T-REX [155], MrBayes [118], and PAML [258].

1.3.1 The principal methods of phylogenetic tree reconstruction

There are four main methods of phylogenetic tree reconstruction. Here, we present an overview of those methods:

Distance methods. Distance methods compute pairwise distances prior to the phylogenetic tree reconstruction. When the pairwise distances are sufficiently close to the actual number of evolutionary changes between species, the distance methods can reconstruct a correct tree. However, in most cases, it is necessary to correct the pairwise distances so that they account for multiple substitutions at the same site. There are several continuous time Markov models for modelling sequence evolution and correcting pairwise distances. The most popular ones are: Jukes Cantor [126], Kimura 2-parameter [132], and Hasegawa-Kishino-Yano (HKY) [105].

Once a distance matrix is obtained, several clustering techniques can be used to infer the phylogenetic tree that correlate well with the given distances. The UPGMA (Unweighted Pair-Group Method using Arithmetic averages) [202] method was originally proposed for taxonomic purposes [157]. It is adequate when the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages. Neighbour-joining [205] is the most popular among the distance methods. Neighbor-Joining, unlike UPGMA, considers unequal rates of evolution on different branches of the tree. There exists several variants to Neighbor Joining such as BioNJ [93], unweighted Neighbor-Joining [92], etc. It is worth noting that there exists several distance-tree transformation

methods, the popular ones being ADDTREE [211], the Method of Weighted least-squares [159], and FITCH [89]. The main advantage of distance methods is their small time complexity that makes them suitable to the analysis of large datasets.

Maximum parsimony. The maximum parsimony methods are the ones that are most commonly used by biologists due to their biological insight and their simplicity. They infer phylogenetic trees by assessing the possible mutations between sequences. In general, the maximum parsimony methods find the phylogenetic tree that have the minimum number of mutations needed for explaining the differences in the observed multiple sequence alignment (see Figure 1–6 for an example).

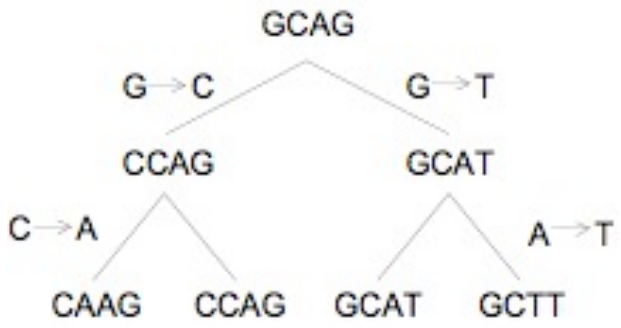


Figure 1–6: The phylogenetic tree with the minimum number of mutations between CAAG, CCAG, GCAT, and GCTT. This figure is taken from [157].

Several variations of parsimony methods exist according to different criteria such as reversibility of nucleotide changes. Methods such as Fitch [90] and Wagner [80] allow reversibilty while Dollo [81] and Camin-Sokal [30] does not. The principle of parsimony has also been generalized to account for different substitution scores among nucleotides [210].

Maximum likelihood. The maximum likelihood methods assign quantitative probabilities to mutational events instead of counting them as done in maximum parsimony. The principle of maximum likelihood in phylogenetics

analysis has been introduced by Felsenstein [82]. Here, all the space of possible phylogenetic trees related to the multiple sequence alignment are assessed according to their ability to predict the observed multiple sequence alignment. The phylogenetic tree with the highest probability of producing the multiple sequence alignment is chosen. Maximum likelihood reconstructs ancestors for all the internal nodes of the tree and computes also the branch lengths according to the mutation probabilities. It is worth noting that for each possible phylogenetic tree, various techniques allow to optimize the likelihood of producing the multiple sequence alignment by varying branch lengths, according to the given probabilistic evolutionary model.

The computational complexity of maximum likelihood computation makes it applicable to only small datasets (less than 100 taxa). However, new heuristics have been developed to handle large data and to be as rapid as parsimony methods. The most used packages, such as DNAML in PHYLIP [85] and PAUP [229] use a hill climbing technique by combining insertion taxa in a growing tree and topological rearrangements [101]. Several enhancements and algorithmic changes done in fastDNAml has improved performance and reduced memory usage, making it feasible for about hundred taxa [184]. However, PHYML currently constitutes one of the best improvement in rapidity and memory requirements [101]. PHYML optimizes a tree topology and branch lengths of a unique tree that is progressively modified such that the tree likelihood increases at each step. It can easily and rapidly be applied to large datasets with more than five hundred taxa with high accuracy [101]. There exists multiple maximum likelihood programs that implement different strategies to break the computation expensive such as NJML that combine the famous Neighbour Joining to maximum likelihood [185], Puzzle that decomposes the phylogenetic tree into quartets [227], and many more [146, 198, 206, 258].

Bayesian methods. Bayesian Methods are an alternative to maximum likelihood due to their time complexity. Here, the methods find the phylogenetic tree that maximizes the posterior probability of the data, which is proportional to the likelihood times the prior probability of that phylogenetic tree. The Bayesian methods have multiple advantages. They generate a number of trees, allow estimation for their posterior probabilities, and they take into account sources of uncertainty that the standard maximum likelihood methods ignores. Moreover, they widely used stochastic optimization and Makov Chain Monte Carlo algorithms while exploring the space of solution [118, 147, 196]. The Bayesian methods are relatively new in phylogenetic analyses and the results provided by these methods are quite accurate [140]. The well-know software MrBayes reconstructs phylogenetic trees for different types of biological and discrete sequences [118]. BAMBE can reconstruct phylogenies only from multiple DNA sequence alignment [218].

1.3.2 Phylogenetic networks

A phyogenetic tree is a standard way of representing the evolution of living organisms. However this representation captures only vertical dependencies between the studied species. In fact, there exists multiple biological mechanisms occurring during the evolution that cannot be depicted by classical phylogenetic tree. The mechanisms of horizontal gene transfer, hybridization, homoplasy and recombination are much more complex to model and are usually presented through a network model [121, 160]. Phylogenetic networks permit the representation of conflicting phylogenetic signals or alternative phylogenetic histories in a single figure. Figures 1–7 and 1–8 present two different types of network models. Figure 1–7 shows a Neighbor Net program [28] representation based on the split tree principle. In this figure, the presence of recombination leads to a complex set of relationships among viral

species of the tomato-infecting begomoviruses [192]. Figure 1–8 presents a result of a horizontal gene transfer detection using T-Rex program [155] for a set of Archaea species.

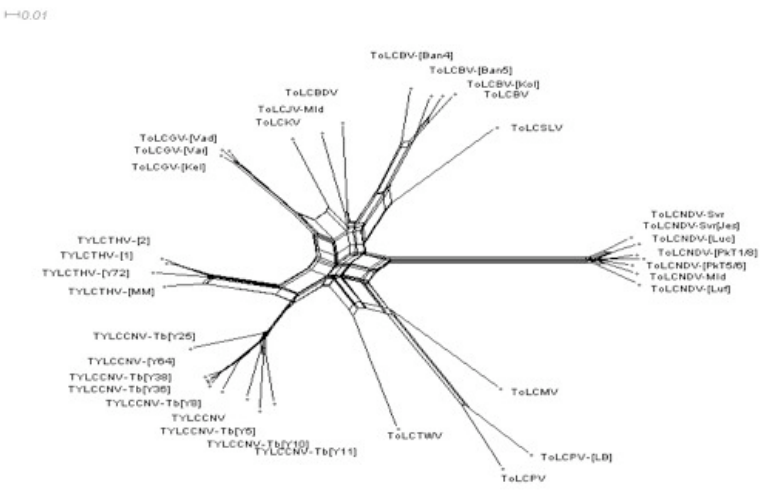


Figure 1–7: Neighbor-Net generated for the tomato-infecting begomoviruses of South and Southeast Asia. Networked relationships among the viral species with boxes, instead of bifurcating evolutionary tree indicate the presence of recombination. This figure is taken from [192].

1.4 Ancestral genomes reconstruction

Since it has been shown that the phylogeny of eutherian mammals is such that an accurate reconstruction of the genome of an early ancestral mammal is possible [16], a lot of interest has been given to the reconstruction procedure. An accurate reconstruction of ancestral genomes will help on various studies such as adaptation, behavioral changes, functional divergences, etc. [135]. However, its reconstruction involves several difficult steps. For more details in this subject, see Appendix A. Here, we briefly describe the different steps of the reconstruction procedure.

The prediction of ancestral genomes can be decomposed into four main steps. A crucial first steps toward the reconstruction is to build an accurate

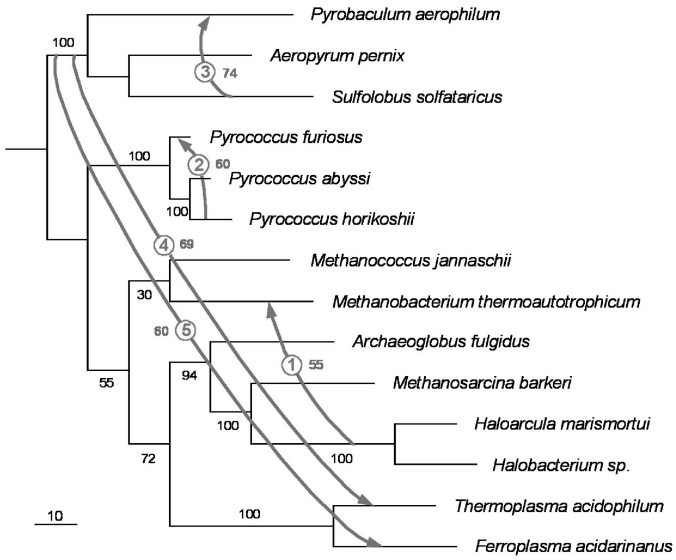


Figure 1–8: Horizontal gene transfer representation with T-Rex for 14 species of Archaea [162]. Numbers on HGT arrows indicate their order of appearance in the unique gene transfer scenario found by the HGT detection method. Bootstrap scores for transfers are indicated by numbers close to arrow circles. Taken from our paper presented in Appendix C [156].

multiple alignment of the extant orthologous sequences, thus establishing orthology relationships among the nucleotides of each sequence. Second, the process of indel reconstruction determines the most likely scenario of insertions and deletions that may have led to the extant sequences. Third, substitution history is reconstructed using a maximum likelihood approach. The last step involves dealing with genome rearrangements (inversions, transpositions, translocations, duplications, and chromosome fusions, fissions, and duplications). One can group the three first step into the problem of reconstructing the ancestral sequences for a set of orthologous region of different organisms. This problem can be defined as follows: given a phylogenetic tree relating the evolutionary history of the organisms, the DNA or amino acids sequences of orthologous regions of the organisms; find the ancestral sequences at each node of the tree (Figure 1–9).

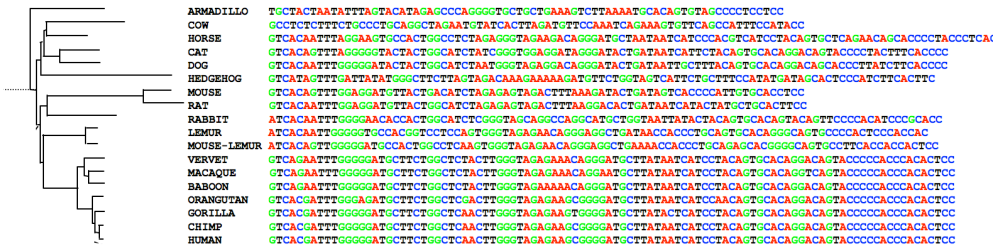


Figure 1–9: The Input of the ancestral sequence reconstruction. This figure is adapted from [16].

1.4.1 Multiple sequence alignment for orthologous sequences for large genomic regions

Here, the multiple sequence alignment uses the evolutionary criterion discussed in Section 1.2.2. Given a set of orthologous sequences, the multiple alignment problem consists of identifying (by aligning them together) the sets of nucleotides derived from a common ancestor through direct inheritance or through substitution. Many approaches have been developed to align multiple, large genomic regions. Some of the most popular approaches include programs like MAVID [25], MLAGAN [27, 42], and TBA [17]. All these approaches fall under the category of progressive alignment methods, and require the prior knowledge of the topology of the phylogenetic tree that relates the extant sequences compared (see Section 1.2.2). The threaded blocks aligner (TBA) program, based on the well-established pair-wise alignment program BLASTZ [212], has been shown to be particularly accurate for aligning mammalian sequences and is thus a tool of choice for ancestral reconstruction for these species. After the alignment procedure, the nucleotide sequences given in Figure 1–9 have been grouped by column according to their predicted common history (Figure 1–10).

1.4.2 Indel reconstruction

Here, we briefly introduce the problem of indel reconstruction. For more details, see the Chapters 2 and 3 dedicated to the indel reconstruction problem.

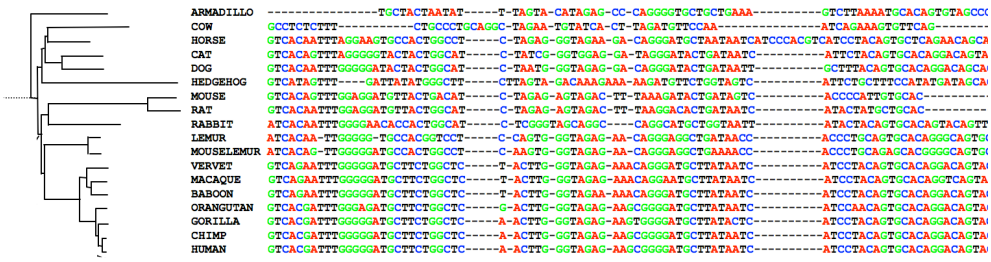


Figure 1–10: The multiple sequence alignment puts in each column only nucleotide that share a common ancestor. The absence of a character in a given position is indicated by a gap (-). This figure is adapted from [16].

After obtaining the multiple sequence alignment of the extant sequences and a phylogenetic tree with known topology and branch lengths, the next step consists of predicting, for each ancestral node in the tree, which columns of the alignment correspond to ancestral bases, and which correspond to nucleotides inserted after the ancestor (Figure 1–11). While the problem of parsimonious indel inference has recently been shown to be NP-Hard [35], good heuristics have been developed by Fredslund et al. [91], Blanchette et al. [16], and Chindelevitch et al. [35]. The maximum likelihood indel problem has been recently adressed by Diallo et al. [62], Kim and Sinha [131], and Bradley and Holmes [23] using respectively phylogenetic-hidden Markov model, hidden Markov model and transducers.

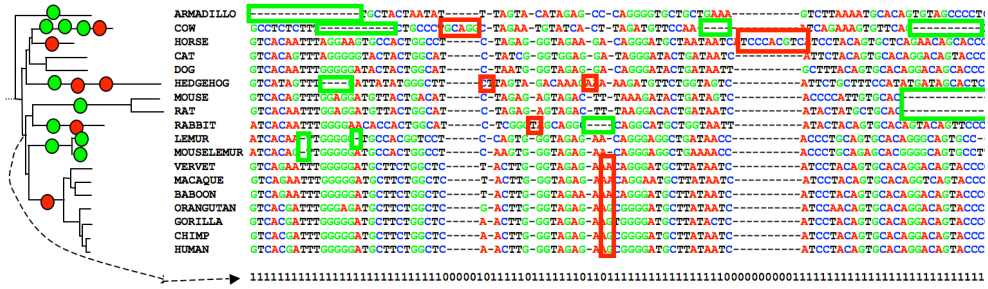


Figure 1–11: The results of the indel reconstruction in which gaps have been mapped as insertion (red square) or deletion (green square). The sequence below presents the content of one ancestor after the indel reconstruction. This figure is adapted from [16].

1.4.3 Substitutions reconstruction

After having established which positions of the multiple alignment correspond to bases in the ancestor, as presented in Figure 1–12, the next step is to predict which nucleotide (A, C, G, or T) was present at each position of a given ancestor using the standard posterior probability approach. The one used by Blanchette et al. in 2004 [16] was based on a dinucleotide substitution model where substitutions at two adjacent positions are independent except for CpG, whose substitution rate to TpG is ten times higher than those of other transitions [259, 215]. This phase of the reconstruction relies on the availability of accurate branch length estimates for the phylogenetic tree.

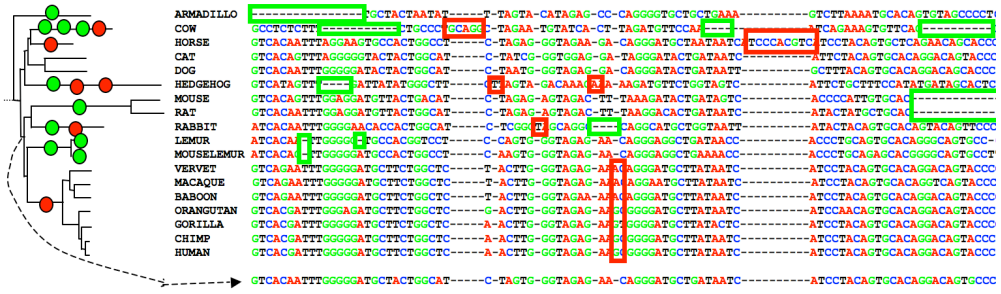


Figure 1–12: The results of the substitution reconstruction identify the nucleotide present in each character of each ancestor.The sequence below presents the content of one ancestor after the substitution reconstruction. This figure is adapted from [16].

1.4.4 Genome rearrangements

To complete the inference of ancestral genomes, the ancestral DNA sequences inferred for each block of orthologous sequences need to be ordered into a single, contiguous genome. This problem is made challenging by the presence of genome rearrangements (inversions, transpositions, translocations, and duplications/losses). One of the most popular computer programs for inferring ancestral gene arrangement is MGR ([21].

1.5 Overview of the thesis

Indel evolutionary scenarios are useful in several problems such as studying evolutionary processes, genome function, adaptation and convergence, annotation of functional regions of extant genomes, including protein-coding regions, RNA genes and others. However, the ancestral genome reconstruction procedure includes several sophisticated steps [15] as mentioned previously. The second and third steps involve the inference of the set of substitutions, insertions and deletions that may have produced a given set of multiply-aligned sequences for a group of extant species. While the problem of reconstructing substitution scenarios has been well-studied [82, 83, 210], the inference of insertion and deletion (indel) scenarios has received less attention (in particular the indel parsimony problem [35, 91]). Recently, the indel maximum likelihood problem has also been adressed during my thesis [131]. The difficulty of the indel reconstruction problem is in large part due to the fact that insertions and deletions often affect several consecutive nucleotides. Thus, the columns of the alignment cannot be treated independently, as opposed to the maximum likelihood problem for substitutions [82].

Given a multiple alignment of orthologous DNA sequences and a phylogenetic tree for these sequences, we have proposed an exact algorithm for the problem of reconstructing the most likely scenario of insertions and deletions capable of explaining the gaps observed in the alignment [62]. We also designed a new statistical framework for indel analysis from a given alignment and a related phylogenetic tree. The new statistical framework provides a way of weighing insertions and deletions of various lengths against each other. Moreover, it provides an accurate probabilistic model of indels, an exact and heuristic algorithm for the reconstruction of indel scenarios, and allows the estimation of the uncertainty for each part of the solution [58]. Similarly to the

statistical alignment approaches [108, 109, 152], which unfortunately remain too slow for genome-wide reconstructions, our method seeks to gain a richer insight into ancestral sequences and evolutionary processes of more than 20 taxa. It is important to notice that the results given by such a reconstruction can lead to several competing solutions that would be necessary to be reported with their confidence level. Hence, we have proposed a way of visualizing the $n-$best indels scenario in ancestral genome reconstruction. We also proposed a colored output of the results corresponding to the ancestral reconstruction confidence. Most of these realisations are present in our Ancestors 1.0 program available at: <http://ancestors.bioinfo.uqam.ca/ancestorWeb/>. It can be integrated into the pipeline of the project of the ancestral mammalian reconstruction initiated by David Haussler from the University of California at Santa Cruz (UCSC), with the collaboration of several other universities such as Pennsylvania State University and McGill University.

Several genomics data studies rely on the availability of either accurate multiple sequence alignments or accurate phylogenetic trees from the data to be studied[1] . Most of the existing approaches for building a multiple sequence alignment and a phylogenetic tree from a given set of sequences first build a mutiple sequence alignment, and then reconstruct the phylogeny based on this mutiple sequence alignment. However, such a direct dependency of mutiple sequence alignment and phylogenetic reconstruction can lead to biased estimations. Thus, the ideal solution is the joint inference of both of them. However, existing methods are limited to small datasets (three to five species)[107]. In this thesis, we presented a useful application of indel score towards the assessment of phylogenetic tree and multiple sequence alignment accuracies. The

[1] In some cases, both accurate alignment and phylogeny are required.

latter score can lead to the use of iterative approaches in the joint inference of phylogenetic tree and multiple sequence alignment.

The design of algorithmic tools in bioinformatics must serve in the analyses of real biological data. Thus, the third part of this thesis focuses on the application of the developed methods in large data analyses. The first application describes to the study of the phages classification. While most of the organisms evolve in a tree-like evolution, the phages have several evolutionary mechanisms such as horizontal gene transfer, duplication and gene losses that cannot not be illustrated using a classical phylogenetic tree. Here, a network is more suitable to represent the evolution of those organisms [103, 104, 144]. One other important problem in the classification of phages is the large range of genomic sizes. Hence, one might consider a special approach of aligning genomes and reconstructing the species tree. Here, we are also interested in reconstructing the ancestral protein sequences to identify the rise of new biological functions. The second dataset analyzed during this thesis is the Human Papilloma virus family and their carcinome classification. Here, we presented the first complete whole genome tree of the Human Papilloma virus. We analyzed indel frequencies according to the carcinome classification. The highest indel frequencies are in the subtrees where there are only low risks of carcinogenicity. Then, we also designed an algorithm intended for finding genomic regions that may be responsible for HPV carcinogenicity. The algorithm is based on the hypothesis that sequence regions responsible for cancer are expected to be very similar among the carcinogenic Human Papilloma Viruses while they should differ a lot from the homolog regions in the non-carcinogenic Human Papilloma Viruses.

Missing data and horizontal gene transfer detection is an important part of all our studies due to the analyzed data. Hence, we finish this thesis by presenting a contribution in both cases.

1.6 Thesis roadmap

This thesis is organized as follows. In Chapter 1 we provided some background on species evolution, genome organization, genomic data analysis through multiple sequence alignment, phylogenetic tree reconstruction, ancestral genome reconstruction and thesis overview and roadmap. A brief review of the recent studies on indel inference is presented in Chapter 2. Chapter 3 presents the inference of indel maximum likelihood scenarios using exact and heuristic approaches. The Chapter 4 focuses on the representation of the results of indel inference. There, we present a web interface implementing the tools developed. We also present an approach for representing the $n-$best competing indel scenarios. The first application, using indel scores for the assessment of multiple sequence alignment and phylogenetic tree, is shown in Chapter 5. Chapter 6 studies phages classification while the Chapter 7 focuses on the evolutionary study of the Human Papilloma Viruses and their carcinome classification. A conclusion of all the thesis is presented in Chapter 8, followed by appendices presenting additional contributions in computational tools for ancestral sequence reconstruction, missing data analyses, horizontal gene transfer detection and description of a dynamic programming algorithm for ancestral profile alignment. To improve the thesis readability, several other appendices[2] are available in the thesis web page <http://ancestors.bioinfo.uqam.ca/phdDiallo/>.

[2] Those appendices contain collection of data, raw results of the related studies, appendices of the published papers, etc.

1.7 Publications and author contributions

This thesis includes a partial or full text and figures of ten scientific articles (including journals, proceedings and book chapter). six articles have been published or accepted for publication. The remaining articles are in preparation. I am the first author and major contributor of eight of the ten papers. I am second author in one paper and last author in one paper. It is worth noting that during my Ph.D., I have written five other articles (four of them have already been published) that are not included in the thesis [53, 54, 61, 95, 181]. I am first author or major contributor of four or these papers. Below is presented the list of the included contributions according to the thesis chapters and my contribution to each paper.

- **Chapter 1**

  This chapter is based in part on:

  - Blanchette, M., **Diallo, A.B.**, Green, E. D., Miller, W. and Haussler, D. (2007): Computational reconstruction of ancestral DNA sequences. Chapter 11 of the book : Methods in Molecular Biology: Phylogenomics. Edited by: W. J. Murphy, Humana Press Inc., Totowa, NJ, 171-184.

    My contribution to this book chapter was the literature review and the chapter redaction. I did the summary of the principal source of information used in this chapter. The materials used in this chapter come principally from a paper published by Blanchette and the other coauthors in 2004 [16].

- **Chapter 2**

  Parts of this chapter come from:

  - **Diallo, A.B.**, Makarenkov, V., and Blanchette, M. (2006): Finding Maximum Likelihood Indel Scenarios. Proceedings of Recomb-CG 2006, Lecture Notes in Computer Science, 4205, Springer Verlag, 171-185.

    All the work in this publication has been done by me under my advisors supervision.

- **Chapter 3**

  This chapter contains the full text of:

  - **Diallo, A.B.**, Makarenkov, V., and Blanchette, M. (2007): Exact and heuristic method to the indels maximum likelihood problem. Journal of Computational Biology. 14 (4), 446-461.

    All the work in this publication has been done by me under my advisors supervision.

- **Chapter 4**

  This chapter contains the text of:

  - **Diallo, A.B.**, Makarenkov, V., Blanchette, M. (2009): Ancestor 1.0: A web interface for ancestral sequence reconstruction. *In preparation.*

    All the work in this publication has been done by me under my advisors supervision.

– **Diallo, A.B.**, Gaul, E. (2009): Visualization of the $n-$best indel likelihood scenarios. 14 pages. *In preparation.*

I contributed equally to this paper with Mr. Eric Gaul. I did the modeling of the indel scenarios handling and the confidence analyses. I designed and implemented the algorithm to find the $n-$best indel scenarios. Mr. Gaul implemented the java library for the visualization.

- **Chapter 5**

This chapter contains the full text of:

– **Diallo, A.B.**, Makarenkov, V., Blanchette, M. (2009): Indel score for the assessment of multiple sequence alignments and phylogenetic trees reconstruction accuracy. *In preparation.*

All the work in this publication has been done by me under my advisors supervision.

- **Chapter 6**

This chapter contains the full text of:

– **Diallo, A.B.**, Nguyen, D., Badescu, D., Boc, A., Blanchette, M. and Makarenkov, V. (2009): Étude de classification des bactériophages. Mathématiques, Informatique et Sciences Humaines. 16 pages. *Submitted.*

I have contributed equally with Dung Nguyen in this paper. We are

joint first authors of the publication. I did the ancestral protein sequence reconstruction and the phylogenetic analyses. I implemented the framework for the ancestral sequences studies. D. Nguyen collected data, did the cognitive analyses and currently manages the database. A. Boc was responsible for the horizontal gene transfer analyses. D. Badescu contributed to the implementation of the rest of the framework. This work was done under the guidance of my advisors.

- **Chapter 7**

  This chapter contains the full text of:

  - **Diallo, A.B.**, Badescu, D., Makarenkov, V., Blanchette, M. (2009): A whole genome study and identification of specific carcinogenic regions of the Human Papilloma Viruses. Journal of Computational Biology. *Accepted for publication.*

    I am the major contributor of this paper. I designed and implemented the algorithms presented in the paper, implemented the p-value computations, and I also did the indel analyses. Badescu D. retrieved, managed and prepared the data and analyzed the prediction of carcinogenicity. This work was done under the guidance of my advisors.

- **Appendix A**

  This chapter contains the full text of:

  - Blanchette, M., **Diallo, A.B.**, Green, E. D., Miller, W. and Haussler, D. (2007): Computational reconstruction of ancestral DNA

sequences. Chapter 11 of the book : Methods in Molecular Biology: Phylogenomics. Edited by: W. J. Murphy, Humana Press Inc., Totowa, NJ, 171-184.

My contribution to this book chapter was the litterature review and the chapter redaction. I did the summary of the principal source of information used in this chapter. The materials used in this chapter come principally from a paper published by Blanchette and coauthors in 2004.

- **Appendix B**

  This chapter contains the full text of:

  - **Diallo, A.B.**, Makarenkov, V., Blanchette, M. and Lapointe, F.-J. (2006): A new efficient method for assessing missing nucleotides in DNA sequences in the framework of a generic evolutionary model. Proceedings of the meeting of the International Federation of Classification Societies 2006, Data Science and Classification. eds Batagelj, V., Bock, H.H., Ferligoj, A., Ziberna, A., Springer Verlag, Ljubljana, 333-340.

    All the work in this publication has been done by me under my advisors supervision.

- **Appendix C**

  This chapter contains the full text of:

  - Makarenkov, V., Boc, A., Diallo Al. Bo. and **Diallo Ab.Ba.** (2008): Algorithms for detecting complete and partial horizontal

gene transfers: Theory and practice, in Data Mining and Mathematical Programming, P.M. Pardalos and P. Hansen eds., CRM Proceedings and AMS Lecture Notes, 45, 159-179.

My contribution in this publication is about 20%. I did the procedure of the simulation studies. V. Makarenkov and A. Boc did all the model, the implementation and the bootstrap validation.

CHAPTER 2
Review of Indel studies

2.1 Preface

This chapter makes a review of recent indel studies. It begins with the presentation of the biological origin of indels and the different kinds of bioinformatics studies related to those events. The second part of this chapter describes the problem of finding the best indel scenario. It briefly introduces the actual approaches used to solve this problem. Different sections of this chapter have been taken from Diallo et al. [62].

2.2 Biological origins of indels

Insertions and deletions, as well as the other types of mutations, can arise in two different contexts [190]:

- chromosome replication, where errors sometimes occur in the replication that lead to changes in the DNA sequences. This phenomenon is called ***replication-dependent mutations;***
- alteration of the DNA sequences in processes that are independent of DNA replication. This is identified as ***replication-independent mutations.*** Transposable elements or transposons, which can move from one site to another without requirement for sequence relatedness at the donnor and acceptor sites, are the most important source of the alteration.

Indel mutations affect both somatic and germline cells. However, due to the fact that only the germline mutations are inherited by descendants, only those mutations could have a consequence in genome evolution and could be

observed in comparative genomics. Replication-dependant indels are generally due to a looping out of either the template DNA strand or the growing strand during the DNA syntesis [190]. Moreover, replication slippage leads to replication errors by the the DNA polymerase, which could lead to the deletion or insertion of DNA segment. The latter phenomenon is more likely to occur between neighbouring repeats. In the case of an insertion, it will correspond to a duplication of the corresponding DNA segment. The other source of replication errors is segments of identical bases in the DNA. These mechanisms often produce short indels (up to 20 to 30 nucleotides). Longer indels are mainly due to either recombination via unequal crossing over, exon shuffling and transposition or transposons. It is worth noting that transposons constitute a large fraction of the eukaryotic genomes size [16, 130]. Finally, it is important to notice that horizontal gene transfers can be viewed as indels if the donor lineages are unknown.

Indels can be divided into three categories:

- indel events occuring in the intronic or intergenic regions, which do not have not a direct impact on protein production;
- mutiple of three nucleotide indels in translated DNA regions, which cause a deletion or an insertion of new amino acids;
- non multiple of three nucleotide indels in translated DNA regions cause shifts in the reading frame. This leads to have changes in most of the produced amino acids and it is also likely to modify the position stop codon [190].

Moreover, a recent study divided indels into five classes (indels of a single base, monomeric base pair expansions/contractions, multi-base pair expansions of 2-15bp repeats units, transposon insertions, and indel containing random sequences) [165]. Table 2–1 gives the percentage of the different classes in the

Human Genome. Ryan and coauthors found that the density of indel in the human genome is about one indel per 7.2kb of DNA, and more than 36 % of the indels are found in known genes and promoters [165]. Several studies pointed out the importance of indels in various diseases such as Huntington's disease, fragile X syndrome, myotonic dystrophy and many types of tumors [70, 190, 232, 240, 249].

Table 2–1: The percentage of the different classes of indels in the human genome. The values presented in this table are extracted from [165].

Indel Class	**% of indels**
Single bases	29.1
Monomeric base pair expansions	18.5
Multibase pair expansions	11
Transposons	0.6
Random sequences	40.8

2.3 Diversity of indels studies

Recently, many studies have focused on indel events. Those studies can be divided into three categories. The first category studies genetic variation within species haplotypes [13, 44, 165, 250]; the second one studies indels in genomes evolution among species [12, 151, 220]; and the third one concerns methods attempting to reconstruct indel scenarios from species evolution [16, 35, 58, 91, 131]. In the next sections, we will focus on the third category. However, we will begin with a brief review of some results obtained by the two first categories. First, an initial map of indel variation in the human genome has been produced with more than 400,000 unique indel polymorphisms and it is available in dbSNP of the NCBI [40, 165, 241]. Second, Lunter and coauthors have designed a neutral indel model for the identification of human functional DNA. They account for indel rate variation and compute the intergap distance distribution between human and mouse. They found that the intergap distance follows a geometric distribution and deviations observed can be related to indel

selective pressure [151]. Moreover, indels in human protein coding regions have been found to be subject to different levels of selective pressure in relation to their structural impact on the amino acid sequence. Furthermore, indels are more likely to be found in regions that do not form important structural domains [45]. In another context, indels are used as phylogenetic markers and have been shown to help improve phylogeny reconstruction [6, 204] as well as multiple sequence alignment reconstruction[134, 150]. Various studies are now trying to treat gaps in a probabilistic framework. However the results of those methods are limited to small numbers of sequences [152, 163, 236, 237].

2.4 The Indel Scenario Problem

In this section, we will give a precise definition for the problem of finding the best indel scenario for a given set of orthologous sequences. Consider a rooted binary phylogenetic tree $T = (V_T, E_T)$ with branch lengths $\lambda : V_T \rightarrow \mathbb{R}^+$. If n is the number of leaves of T, there are $n-1$ internal nodes and $2n-2$ edges [1] .

Consider a multiple sequence alignment A of n orthologous sequences corresponding to the leaves of the tree T. Since the only evolutionary events of interest here are insertions and deletions, A can be transformed into a binary matrix, where gaps are replaced by 0's and nucleotides by 1's. Let A_x be the row of the binarized alignment corresponding to the sequence at leaf x of T, and let $A_x[i]$ be the binary character at the i-th position of A_x. Assuming that the alignment A contains L columns, we add for convenience two extra columns, $A[0]$ and $A[L+1]$, consisting exclusively of 1's.

[1] It is worth noting that certain methods can deal with unrooted trees (for instance see [91])

Definition 2.1 (Ancestral reconstruction) *Given a multiple alignment A of n extant sequences assigned to the leaves of a tree T, an* ancestral reconstruction *A^* is an extension of A that assigns a sequence $A^*_u \in \{0,1\}^{L+2}$ to each node u of T, and where $A^*_u = A_u$ whenever u is a leaf.*

The following restriction on the set of possible ancestral reconstructions is necessary in some contexts.

Definition 2.2 (Phylogenetically correct ancestral reconstruction) *An ancestral reconstruction A^* is* phylogenetically correct *if, for any $u, v, w \in V_T$ such that w is located on the path between u and v in T, we have $(A^*_u[i] = A^*_v[i] = 1) \implies (A^*_w[i] = 1)$.*

Requiring an ancestral reconstruction to be phylogenetically correct corresponds to assuming that any two nucleotides that are aligned in A have to be derived from a common ancestor, and thus that all the ancestral nodes between them have to have been a nucleotide. This prohibits aligned nucleotides to be the result of two independent insertions. Assuming that this property holds perfectly for a given alignment A is somewhat unrealistic, but, for mammalian sequences, good alignment heuristics have been developed (e.g. TBA [17], MAVID [25], MLAGAN [27]) and have been shown to be quite accurate [17]. It might be necessary, in the future, to relax this assumption, but, for now, we will concentrate only on finding phylogenetically correct ancestral reconstructions.

Since we are considering insertions and deletions affecting several consecutive characters, we delimit each operation by the positions s and e in the aligned sequences where it starts and ends. Let x and y be two nodes of the tree, where x is the parent of y. The pairwise alignment consisting of rows A^*_x and A^*_y is divided into a set of regions defined as follows (see Figure 2.4).

Definition 2.3 (Deletions, Insertions, Conservations, and Length) *Consider the pairwise alignment of A_x^* and A_y^*, and let $0 \leq s \leq e \leq L+1$.*

- *The region (s,e) is a* deletion *if (a) for all $i \in \{s,\ldots,e\}, A_y^*[i] = 0$, (b) $A_x^*[s] = A_x^*[e] = 1$, and (c) no region $(s',e') \supset (s,e)$ is a deletion (i.e. we only consider regions that are maximal).*
- *The region (s,e) is an* insertion *if (a) for all $i \in \{s,\ldots,e\}, A_x^*[i] = 0$, (b) $A_y^*[s] = A_y^*[e] = 1$, and (c) no region $(s',e') \supset (s,e)$ is an insertion.*
- *The region (s,e) is a conservation if (a) for all $i \in \{s,\ldots,e\}, A_x^*[i] = A_y^*[i]$ and (b) no region $(s',e') \supset (s,e)$ is a conservation.*
- *The* length *of region (s,e) is the number of non-trivial positions it contains: $l(s,e) = |\{s \leq i \leq e | A_x^*[i] \neq 0 \text{ or } A_y^*[i] \neq 0\}|$.*

A pair of binary alignment rows A_x^* and A_y^* can thus be partitioned into a set of non-overlapping insertions, deletions, and conservations.

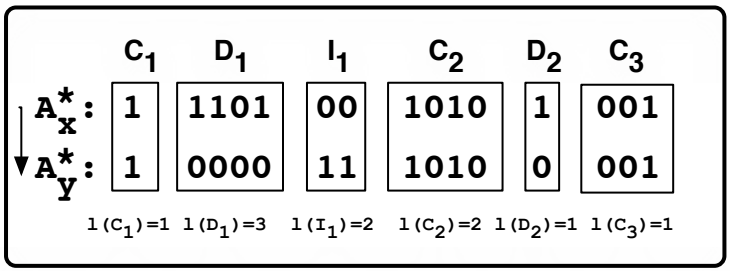


Figure 2–1: Example of the partition of a pairwise alignment of A_x^* and A_y^* (where x is the parent of y) into deletions, insertions, and conservations. The length of each operation is given below it.

Definition 2.4 (Indel scenario) *The indel scenario defined by an ancestral reconstruction A^* is the set of insertions and deletions that occurred between the ancestral reconstructions at adjacent nodes in T.*

All that remains is to define an optimization criterion on A^*. Two main choices are possible: a parsimony criterion or a likelihood criterion. Hence, the next two sections present the Indel Parsimony Problem (*IPP*) and the Indel Maximum Likelihood Problem (*IMLP*).

2.5 The Indel Parsimony Problem (*IPP)*

The parsimony approach for the indel reconstruction problem has been introduced by Fredslund *et al.* ([91]) and Blanchette *et al.* [16]. In its simplest version, it attempts to find the phylogenetically correct ancestral reconstruction A^* that minimizes the total number of insertions and deletions defined by A^*:

$$indelParsimony(A^*) = \sum_{u,v:(u,v)\in E_T} |\{(s,e) : (s,e) \text{ is an indel from } A^*_u \text{ to } A^*_v\}| \tag{2.1}$$

The Indel Parsimony Problem is NP-Hard [35]. Most authors have studied a weighted version of the *IPP* where the cost of indels depends linearly on their length (affine gap penalty). It is worth noting that the existing methods for solving the *IPP* do not consider the edge lengths as input for finding the best scenario.

2.5.1 Algorithm of Fredslund et al.

In their article, Fredslund and coauthors solved the Indel Parsimony Problem by processing a so-called gap graph through heuristics [91]. In their algorithm, the tree $T = (V_T, E_T)$ is unrooted and the branch lengths are ignored. Their algorithm proceeds in three steps:

Step 1: Constructing the gap graph

The gap graph is built from a set of *gap intervals* and the corresponding *tree covering*. The gap intervals correspond to the regions with consecutive identical alignment columns (see Figure 2–2). For each gap interval, the tree covering corresponds to the set of subtrees with the same gap structure. Furthermore, each subtree cannot be extended without including a taxa with a nucleotide. Hence, the tree coverings of the example shown in Figure 2–2 are $F_1 = \{\{2\}\}$, $F_2 = \{\{2\}, \{4,5\}\}$, $F_3 = \{\{1,2,3,4\}\}$ and $F_4 = \{\{1,2\}, \{4\}\}$.

From the tree coverings, a gap graph is induced according to the following rules. Each subtree in the tree coverings corresponds to a vertex in the gap graph. Using consecutive intervals, the edges are induced according to the different comparisons between the subtrees from the tree coverings of all the consecutive intervals. For instance, an edge between subtrees x and y from consecutive intervals is induced according to the following rules :

- Merge two vertices if they cover the same set of leaves (i.e identical subtrees).
- Create a directed edge from x to y, if y includes all the taxa in x.
- Create a directed edge from y to x, if x includes all the taxa in y.
- Create an undirected zigzag edge between x and y, if the two sets share common taxa but none of them includes all the taxa of the other.
- Finally, there will be no edge between x and y, if the two sets share no common taxa.

The Figure 2–3 shows the gap graph obtained from the example presented in Figure 2–2.

Step 2: Preprocessing the gap graph

Once the gap graph is obtained, the algorithm finds the most parsimonious set

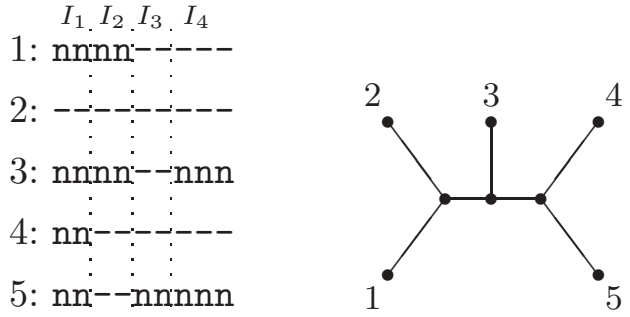


Figure 2–2: Alignment with five sequences and four gaps intervals and the related phylogenetic tree. n represents a presence of nucleotide. This figure is taken from [91].

of indels that explain the alignment, given a gap penalty function $g(l) = \alpha + \beta l$ where l is the length of the gap. To achieve this goal,in this second step, the algorithm attempts to reduce the potentially very large and complex gap graph. Several lemmas and two theorems are required for the latter issue. In summary, the theorems allow a connection between indels and gap graph vertices and suggest to avoid placing indels lower than necessary in the phylogenetic tree. The processing goes over the gap graph in several passes. In each pass, local application of lemmas and theorems allow to reduce the graph. The Figure 2–4 presents an example of reduced graph using the same tree of Figure 2–3. The reduced graph is divided into chains where two vertices belong to the same chain if and only if there exists a path connecting the two vertices that does not cross a leaf vertex. The chains can be treated independently since indels causing gap in the vertices of one chain could not have caused gaps in vertices of another chain.

Step 3: Resolving the reduced gap graph

Each chain is analyzed independently. For instance, the chain a in Figure 2–4 has two indel scenario explanations. 1) There is only one indel occurring between the root of subtree $\{1, 2, 3\}$ and its closest neighbor; 2) The indel in this chain is the result of three independent indels in each leaf branch of the subtree.

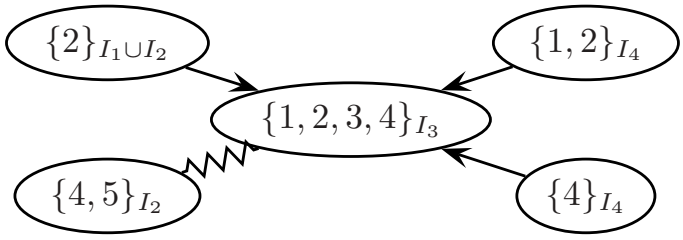


Figure 2–3: The gap graph induced from the example of Figure 2–2. This figure is taken from [91].

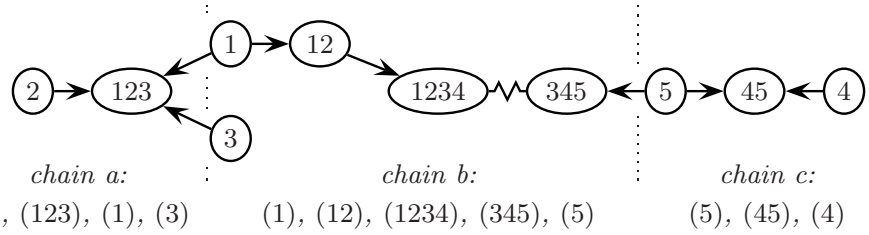


Figure 2–4: An example of reduce gap graph with three distinct chains named a, b and c (same tree as in Figure 2–2). This figure is taken from [91].

This algorithm is exponential in the length of the alignment. However, dependent on the gap configuration in the alignment, the algorithm could be extremely fast (less than one second for an alignment of nine HIV whole genome with about 10000 columns).

2.5.2 Algorithm of Blanchette et al.

The heuristics proposed by Blanchette and coauthors used a greedy approach implemented in the inferAncestor program. The latter program is available from *http://www.mcb.mcgill.ca/˜blanchem/software*. The inferAncestor program integrates the steps of indel and substitution inference. This program is presented in details in the appendix A. Given a multiple alignment, first, all the gaps in the alignment are marked as unexplained. Then, the algorithm iteratively selects the insertion or deletion, performed along a specific edge of the tree and spanning one or more columns of the alignment, that yields the largest number of alignment gaps explained per unit of cost. The number of gaps explained by a deletion is the number of unexplained gaps in the subtree below which the deletion occurs. The number of gaps explained by an insertion is the number of unexplained gaps in the complement of the subtree above which the insertion occurs. The costs can be fixed or defined heuristically. Once the best insertion or deletion has been identified, its gaps are

marked as explained. This does not preclude them from being part of other indels, but they will not count in their evaluation. Finally, heuristics are used to reduce errors due to incorrect alignment, in particular to reduce the problems caused by two repetitive regions from two distantly related species mistakenly aligned to each other, with other species having gaps in that region.

2.5.3 Algorithm of Chindelevitch et al.

In their paper, Chindelevitch and coauthors presented two ways of solving the IPP [35]. The first one is a greedy algorithm and the second one is based on the Integer Linear Programming (ILP). Their simulations indicated that the ILP approach finds the optimal indel score in all simulations while the greedy one gives a nearly optimal score (the score differences are less than 2 in all cases). However these results are based on small datasets. The greedy algorithm works in two phases requiring the application of six different rules. **Phase 1:** the presence or absence of character is identified for each ancestral position in each ancestral node where all the children of the given node at the same position have a common character. **Phase 2:** the heuristic traverses each obtained sequence from the internal nodes and applies one of its rule to decide whether an undecided character should be explained according to the neighboring columns. The process is repeated until all the ancestral characters are inferred. The running time of this greedy algorithm is linear with respect to the number of taxa and the length of the alignment. In the second approach for solving the IPP using ILP, the authors showed an efficient way of encoding the IPP as a 0-1 Integer Linear Programming problem. Once the encoding is done, they used a standard package for linear programming to find the solution. It is worth noting that the latter approach displays an exponential running time in function of the length of the alignment.

2.6 Indel Maximum Likelihood Problem

In this section, we introduce our own version of the indel reconstruction problem in a probabilistic framework similar to the Thorne-Kishino-Felsenstein model [237]. To this end, we need to define the probability of transition between an alignment row A_x^* and its descendant row A_y^*. This probability will be defined as a function of the probability of the insertions, deletions, and conservations that happened from A_x^* to A_y^*.

Let $P_{DelStart}(\lambda(b))$ be the probability that a deletion starts at a given position in the sequence, along a branch b of length $\lambda(b)$, and let $P_{InsStart}(\lambda(b))$ be defined similarly for an insertion. We assume that these probabilities only depend on the length $\lambda(b)$ of the branch b along which they occur, but not on the position where the indel occurs. A reasonable choice is $P_{DelStart}(\lambda(b)) = 1 - e^{-\psi_D \lambda(b)}$ and $P_{InsStart}(\lambda(b)) = 1 - e^{-\psi_I \lambda(b)}$, for some deletion and insertion rate parameters ψ_D and ψ_I, but our algorithm allows for any other choice of these probabilities. Thus, the probability that none of the two events happens at a given position, which we call the probability of a conservation, is given by $P_{Cons}(\lambda(b)) = e^{-(\psi_D+\psi_I)\lambda(b)}$. We make the standard simplifying assumption that the length of a deletion follows a geometric distribution, where the probability of a deletion of length k is $\alpha_D^{k-1}(1-\alpha_D)$ and the probability of an insertion of length k is $\alpha_I^{k-1}(1-\alpha_I)$. One can thus see α_D (resp. α_I) as the probability of extending a deletion (resp. insertion). This assumption, necessary to design a fast algorithm, holds relatively well for short indels, but fails for longer ones [130]. Our algorithm allows the parameters α_D and α_I to depend on the branch b, but the results reported in Section 3.5 correspond to the case where α_D and α_I were held constant across the tree. The probability that alignment row A_x^* was transformed into alignment row A_y^* along branch b can be defined as follows:

$$\Pr(A_y^*|A_x^*, b) = \prod_{(s,e)\text{: deletion from } A_x^* \text{ to } A_y^*} P_{DelStart}(\lambda(b)) \cdot (\alpha_D^{l(s,e)-1}(1-\alpha_D)) \cdot \prod_{(s,e)\text{: insertion from } A_x^* \text{ to } A_y^*} P_{InsStart}(\lambda(b)) \cdot (\alpha_I^{l(s,e)-1}(1-\alpha_I)) \cdot \prod_{(s,e)\text{:conservation from } A_x^* \text{ to } A_y^*} (P_{Cons}(\lambda(b))^{l(s,e)} \quad (2.2)$$

This allows us to formulate precisely the problem addressed in this paper:

INDEL MAXIMUM LIKELIHOOD PROBLEM (IMLP):

Given: A multiple sequence alignment A of n orthologous sequences related by a phylogenetic tree T with branch lengths λ, a probability model for insertions and deletions specifying the values of ψ_D, ψ_I, α_D, and α_I.

Find: A maximum likelihood phylogenetically correct ancestral reconstruction A^* for A, where the likelihood of A^* is:

$$L(A^*) = \prod_{b=(x,y)\in E_T} \Pr(A_y^*|A_x^*, b) \quad (2.3)$$

During my Ph.D., Kim and Sinha published the indelign program for solving the IMLP [131]. Their algorithm is based on pair hidden Markov Model. Recently, Bradley and Holmes improved on the implementation of Indelign using transducers [23]. In the next section we introduce the indelign program and show the differences between the two implementations.

2.6.1 Algorithm of Indelign

Given a multiple sequence alignment, the Indelign program [131] can be used to annotate the indels on each branch of a phylogenetic tree by solving the IMLP, it can also make limited changes to the alignment to infer a better evolutionary history. Furthermore, the Indelign program can infer indel evolutionary paramaters from related multiple sequence alignments and a phylogenetic tree. To solve the IMLP, Indelign builds a pair-hidden Markov

model in each edge of the tree (for example see Figure 2–5). The **ANNOTATE** component of Indelign solves the IMLP either by naively evaluating all the possible indel scenarios or by building a dynamic programming matrix. The dynamic program first builds blocks that can be annotated independently of the other blocks and then resolves the adjacent and indel dependent blocks. Each node u of the tree is associated to a matrix of size 2^k where k corresponds to the number of blocks. This matrix records the likelihood of the maximum likelihood indel scenarios for the subtree rooted u. The time complexity of the naive algorithm is $O(2^{k(n-1))}n$ while that of the dynamic program is $O(2^k n)$, where n is the number of taxa. Hence, in the worst case (blocks correspond to alignment columns) the two methods are exponential in function of the alignment length. The **SEARCH** component of the Indelign program permits to obtain all alignments derived from the given one with a limited class of modifications. Those alignments can be quickly evaluated to help improving the alignment quality with respect to the maximum likelihood evolutionary scenario (including indels and substitutions). Indelign's complexity makes it applicable to only dataset with small number of taxa and short sequence lengths. It also prevents having certain configurations of indel scenario such as having an indel falling in the middle of another indel. It also avoids multiple hits of indels in the same branch. The indelign program is described in details in [131].

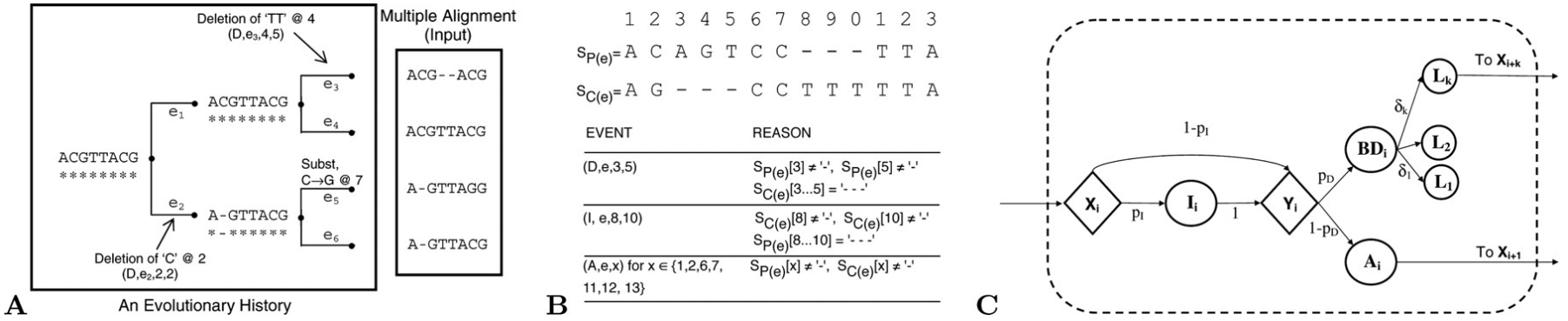


Figure 2–5: A) An example of an evolutionary scenario of Indelign; B) Detailed analysis of a single edge, with the different events; C) Part of the Pair-hidden Markov model for evaluating the maximum likelikood scenarios in the branch. The scenario includes substitutions, however they can be treated independently. The model is based on the parent trimmed sequence Sp(e). Each column i of the trimmed parent sequence has modeled and extra states are added: begin deletion (BD) state, an align (A) state and Insertion (I) state. The values on the edges correspond to the transition probabilities. This figure is taken from [131].

2.6.2 Algorithm of Bradley and Holmes

Recently, Bradley and Holmes showed how the Indelign program can be efficiently implemented using finite-state transducers instead of the pair-hidden Markov model [23]. A transducer is a machine similar to a pair-hidden Markov model: it is a two-tape finite state machine with transition and emission weights [67]. Their method places a transducer on each branch of the phylogenetic tree and automates the construction of systematic scoring schemes for solving the IMLP. The complexity of the transducer approach is $O(la^{2n})$ where a is the number of state in the transducer ($a = 3$ in the simplest transducer). To make this approach practical for a realistic dataset, a Markov Chain Monte Carlo (MCMC) approach has been implemented to sample from the posterior distribution over indel scenarios. It starts with an initial estimate of indel scenario and then computes successive local MCMC moves (edge and node sampling) [23]. It is worth noting that the authors do not make available an implementation.

CHAPTER 3
Exact and Heuristic Methods for the Indels Maximum Likelihood Problem

3.1 Preface

In this chapter, we present a solution to the Indel Maximum Likelihood Problem (IMLP) described in chapter 3. We proposed to solve the IMLP using a new type of hidden Markov model called tree-HMM. The content of this chapter is taken from the published paper Diallo et al. [58] (see appendices for the full version of this paper).

3.2 Abstract

Given a multiple alignment of orthologous DNA sequences and a phylogenetic tree for these sequences, we investigate the problem of reconstructing the most likely scenario of insertions and deletions capable of explaining the gaps observed in the alignment. This problem, that we called the Indel Maximum Likelihood Problem (IMLP), is an important step toward the reconstruction of ancestral genomics sequences, and is important for studying evolutionary processes, genome function, adaptation and convergence. We solve the IMLP using a new type of tree hidden Markov model whose states correspond to single-base evolutionary scenarios and where transitions model dependencies between neighboring columns. The standard Viterbi and Forward-backward algorithms are optimized to produce the most likely ancestral reconstruction and to compute the level of confidence associated to specific regions of the reconstruction. A heuristic is presented to make the method practical for large data sets, while retaining an extremely high degree of accuracy. The methods are illustrated on a 1Mb alignment of the CFTR regions from 12 mammals.

3.3 A Tree-Hidden Markov Model

In this section, we describe the tree hidden Markov model that is used to solve the IMLP. A tree-hidden Markov model (tree-HMM) is a probabilistic model that allows two processes to occur, one in time (related to the sequence history in a given column of A), and one in space (related to the changes toward the neighboring columns). Tree HMMs were introduced by Felsenstein and Churchill (1996) [88] and Yang (1996) [256] to improve the phylogenetic models that allows for variation among sites in the rate of substitution, and have since then been used for several other purposes (e.g. detecting conserved regions [217] and predicting genes [215]). Just as any standard HMM [67], a tree-HMM is defined by three components: the set of states, the set of emission probabilities, and the set of transition probabilities.

3.3.1 States

Intuitively, each state corresponds to a different single-column indel scenario (although additional complications are described below). Given a rooted binary tree $T = (V_T, E_T)$ with n leaves, each state corresponds to a different labeling of the edges E_T with one of three possible events: I (for insertion), D (for deletion), or C (for conservation). The set $\mathcal{S}$ of possible states of the HMM would then be $\mathcal{S} = \{I, D, C\}^{2n-2}$. However, this definition is not sufficient to model certain biological situations (see Figure 3–1). We will use the '*' symbol to indicate that, along a certain branch $b = (x, y)$, no event happened because there was a base neither at node x nor at node y. This will happen in two situations: when edge b is a descendant of edge b' that was labeled with D (i.e. the base was deleted higher up the tree), and when there exists an edge b' that is not between b and the root and that is labeled with I (i.e. an insertion happened elsewhere in the tree). The fact that these extraneous events can potentially interrupt ongoing events along branch b means that the

HMM needs to have a way to remember what event was actually going on along that branch. This transmission of memory from column to column is achieved by three special labels: I^*, D^*, and C^*, depending on whether the * regions is interrupting an insertion, deletion, or conservation. Thus, we have $\mathcal{S} \subseteq \{I, D, C, I^*, D^*, C^*\}^{2n-2}$. Although this state space appears prohibitively large (6^{2n-2}), the reality is that a number of these states cannot represent actual indel scenarios, and can thus be ignored. The following set of rules specify what states are valid.

Definition 3.1 (Valid states) *Given a tree $T = (V_T, E_T)$, a state s assigning a label $s(b) \in \{I, D, C, I^*, D^*, C^*\}$ to each branch $b \in E_T$ is* valid *if the two following conditions hold.*

- *(Phylogenetic correctness condition) There must be at most one branch b such that $s(b) = I$.*
- *(Star condition) Let $b \in E_T$, and let $anc(b) \subset E_T$ be the set of branches on the path from the root to b. Then $s(b) \in \{I^*, D^*, C^*\}$ if and only if $\exists b' \in anc(b)$ such that $s(b') = D$ or $\exists b' \in (E_T \setminus anc(b))$ such that $s(b') = I$.*

The number of valid states on a complete balanced phylogenetic tree with n leaves is $O(n \cdot 3^{2n})$ (the number is dominated by states that have a 'I' on a branch leading to a leaf, which leaves all other $2n-3$ edges free to be labeled with either C^*, D^*, or I^*). Although this number remains exponential, it is significantly better than the 6^{2n-2} valid and invalid states.

3.3.2 Emission probabilities

In an HMM, each state emits one symbol, according a certain emission probability distribution. In our tree-HMMs, each state emits a collection of symbols, corresponding to the set of characters obtained at the leaves of T when indel scenario s occurs. Intuitively, we can think of a state as emitting

an alignment column. The following definition formalizes this.

Definition 3.2 *Let s be a valid state for tree $T = (V_T, E_T)$ with root r. Then, we define the output of state s as a function $O_s : V_T \rightarrow \{0, 1\}$ with the following recursive properties:*

1. $$O_s(root) = \begin{cases} 0, & \text{if } \exists x \in V_T \text{ such that } s(x) = I \\ 1, & \text{otherwise} \end{cases} \tag{3.1}$$

2. *Let $e = (x, y) \in E_T$, with x being the parent of y. Then,*

$$O_s(y) = \begin{cases} 0, & \text{if } s(e) = D \\ 1, & \text{if } s(e) = I \\ O_s(x), & \text{otherwise} \end{cases} \tag{3.2}$$

Let C be an alignment column (i.e. an assignment of 0 or 1 to each leaf in T). We then have the following degenerate emission probability for state s:

$$Pr_e(C|s) = \begin{cases} 1 \text{ if } O_s(x) = C(x) \text{ for all } x \in leaves(T) \\ 0 \text{ otherwise} \end{cases} \tag{3.3}$$

Thus, each state s can emit a single alignment column C. However, many different states can emit the same column.

Missing data

In presence of missing characters among the input sequences, the emission probability can be adapted such that the equality between $O_s(x)$ and $C(x)$ is assessed according to 0's and 1's in $C(x)$ only. It is worth noting that missing characters are different to gaps noted by $-$. Hence, the presence of missing data increases the number of states for a given column.

3.3.3 Transition probabilities

The last component to be defined is the set of transition probabilities of the tree-HMM. The probability of transition from state s to state s', $\Pr_t(s'|s)$, is a function of the set of events that occurred along the edges of T. Intuitively, $\Pr_t(s'|s)$ describes the probability of the single-column indel scenario s', given that scenario s occurred at the previous column. This transition probability is a function of insertions and deletions that started between the two columns, of those that were extended going from one column to the next. Specifically, we have:

$$\Pr_t(s'|s) = \prod_{b \in E_T} \rho(s'(e)|s(e), b), \tag{3.4}$$

where ρ is given in Table 3–1.

Table 3–1: Edge transition table $\rho(s'(e)|s(e), b)$. Notice that ρ is not a transition probability matrix, since its rows sum to more than one.

$s(e) \setminus s'(e)$	C	D	I	C^*	D^*	I^*
C	$P_{Cons}(\lambda(b))$	$P_{DelStart}(\lambda(b))$	$P_{InsStart}(\lambda(b))$	1	0	0
D	$(1-\alpha_D)P_{Cons}(\lambda(b))$	α_D	$(1-\alpha_D)P_{InsStart}(\lambda(b))$	0	1	0
I	$(1-\alpha_I)P_{Cons}(\lambda(b))$	$(1-\alpha_I)P_{DelStart}(\lambda(b))$	α_I	0	0	1
C^*	$P_{Cons}(\lambda(b))$	$P_{DelStart}(\lambda(b))$	$P_{InsStart}(\lambda(b))$	1	0	0
D^*	$(1-\alpha_D)P_{Cons}(\lambda(b))$	α_D	$(1-\alpha_D)P_{InsStart}(\lambda(b))$	0	1	0
I^*	$(1-\alpha_I)P_{Cons}(\lambda(b))$	$(1-\alpha_I)P_{DelStart}(\lambda(b))$	α_I	0	0	1

3.4 Tree-HMM paths, ancestral reconstruction and assessing uncertainty

We now show how the tree-HMM described above allows us to solve the IMLP. Consider a multiple alignment A of length L on a tree T. A path π in the tree-HMM is a sequence of states $\pi = \pi_0, \pi_1, ..., \pi_L, \pi_{L+1}$. Based on standard HMM theory, we get:

$$\Pr(\pi, A) = \Pr(\pi_0, A_0) \prod_{i=1}^{L+1} \Pr_e(A[i]|\pi_i) \cdot \Pr_t(\pi_i|\pi_{i-1}) \tag{3.5}$$

Figure 3–1 gives an example of an alignment with some of the non-zero probability paths associated.

Theorem 3.1 *Consider an alignment A on tree T. Then $\pi^* = argmax_\pi \Pr(\pi, A)$ yields the most likely indel scenario for A, and a maximum likelihood ancestral reconstruction A^* is obtained by setting $A^*_u[i] = O_{\pi^*_i}(u)$.*

Proof 3.1 *It is simple to show that for any ancestral reconstruction $\hat{A}$ for A, we have $L(\hat{A}) = \Pr(\pi, A)$, where π is the path corresponding to $\hat{A}$. Thus, maximizing $\Pr(\pi, A)$ maximizes $L(\hat{A})$.*

3.4.1 Computing the most likely path

To compute the most likely path π^* through a tree-HMM, we adapted the standard Viterbi dynamic programming algorithm [244]. Let $X(i, k)$ be the joint likelihood of the most probable path ending at state k for the i first columns of the alignment. Let $c \in \mathcal{S}$ be the state made of C's on all edges of T. Since the dummy column $A[0]$ consists exclusively of 1's, c is the only possible initial state. For any i between 0 and $L + 1$ and for any valid state $s \in \mathcal{S}$, we can compute $X(i, s)$ as follows:

$$X(i,s) = \begin{cases} 1, & \text{if } i = 0 \text{ and } s = c \\ 0, & \text{if } i = 0 \text{ and } s \neq c \\ \Pr_e(A[i]|s) \cdot \max_{s' \in \mathcal{S}}(X(i-1, s') \cdot \Pr_t(s|s')), & \text{if } i > 0 \end{cases} \tag{3.6}$$

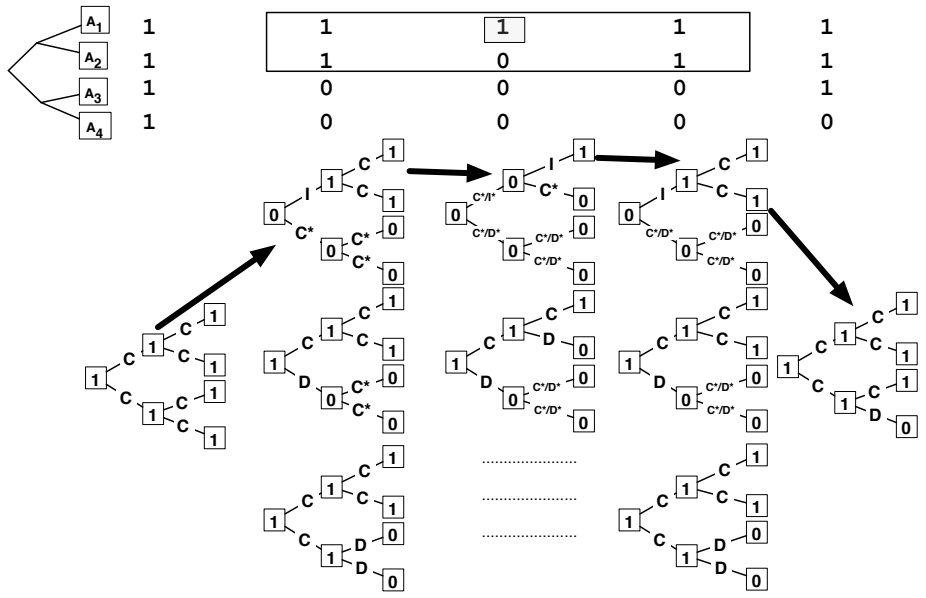


Figure 3–1: The set of valid, non-zero probability states associated to the multiple alignment given at the top of the figure. When edges are labeled with more than one character (e.g. C^*, D^*), the tree represents several possible states. For the third column, not all possible states are shown. Arrows indicate one possible path through the tree-HMM. This path corresponds to two interleaved insertions, shown by two boxes in the alignment, illustrating the need for the I^* character.

Finally, π^* is obtained by tracing back the dynamic programming, starting from entry $X(L+1, c)$. To ensure numerical stability, we use a log transformation and scaling of probabilities as described by [67].

The running time of a naive implementation of the Viterbi algorithm is $O(|\mathcal{S}|^2 L)$, which quickly becomes impractical as the size of the tree T grows. However, we can make this computation practical for moderately large trees and for long sequences. Even though the number of states is exponential in the number of sequences, most alignment columns can only be generated with non-zero probability by a much more manageable number of states. Given an alignment A, it is possible to compute, for each column $A[i]$, the set S_i of valid states that can emit $A[i]$ with non-zero probability. For instance, an alignment column with only 1's will lead to only one possible state, independently of the

number taxa of n. The set S_i can be constructed using a bottom approach presented in Algorithm 3. More states can be discarded by using the fact that the transition probability between most pairs of states is zero. We can thus remove from S_i any state s that is such that the transition to s from any state in S_{i-1} has probability zero. Proceeding from left to right, we get $S'_0 = S_0$, and $S'_i = \{s \in S_i | \exists t \in S'_{i-1} \text{ s.t. } \Pr_t(s|t) > 0\}$, where $S'_i \subseteq S_i$. For instance, if, in all states of S_{i-1}, an edge e is labeled by deletion D, then none of the states in S_i can have edge e labeled with C^* or I^*. This yields a large improvement for alignment regions consisting of a number of adjacent positions with a base in only one of the n species and ensures that the algorithm will be practical for relatively large number of sequences (see Section 3.5).

Algorithm 1 buildValidState(node *root*, C)

Require: *root*: a tree node, C: an alignment column.
Ensure: Set of valid, non-zero probability states for C.

```
1: if root is a leaf then
2:     return list of possible operations according to the character at that leaf
3: else
4:     leftList = buildValidState(root.left, C)
5:     rightList = buildValidState(root.right, C)
6:     return mergeSubtrees(leftList, rightList, root)
7: end if
```

3.4.2 Assessing uncertainties of the ancestral reconstruction

A significant advantage of the likelihood approach over the parsimony approach is that it allows evaluating the uncertainty related to certain aspects of the reconstruction. For example, it is useful to be able to compute the probability that a base was present at a given position i of a given ancestral node u:

$$\Pr(A^*_u[i] = 1|A) = \sum_{s \in \mathcal{S}: O_s(u)=1} \Pr(\pi_i = s|A). \tag{3.7}$$

Algorithm 2 mergeSubtrees(StateList $leftList$, StateList $rightList$, node $root$)

Require: $leftList$ and $rightList$: the lists of partial states, $root$: a tree node.
Ensure: Set of valid, non-zero probability states combining elements in $leftList$ and $rightList$.

```
1: mergedList ← emptyList
2: for all partial states l in leftList do
3:    for all partial states r in rightList do
4:       if compatible(l, r) == true  then
5:          m = merge(l, r)
6:          if root == initialroot then
7:             mergedList.add(m)
8:          else
9:             for op ∈ {C, D, I, C*, D*, I*} do
10:               if isPossibleUpstream(m,op) then
11:                  mergedList.add(addAncestorBranch(m,op))
12:               end if
13:            end for
14:         end if
15:      end if
16:   end for
17: end for
18: return mergedList
```

This allows the computation of the probability of making an incorrect prediction at a given position of a given ancestor. The forward-backward is a standard HMM algorithm to compute $\Pr(\pi_i = s|A)$ (see [67] for more details). The optimizations developed for the Viterbi algorithm can be trivially adapted to the Forward-Backward algorithm.

3.5 Results of the exact method

Our tree-HMM algorithm was implemented as a C program that is available upon request. The program was applied to a ∼700kb region of the CFTR locus on chromosome 7 of human, together with orthologous regions in 11 other species of mammals: chimp, macaque, baboon, mouse, rat, rabbit, cow,

dog, Rodrigues fruit bat (rfbat), armadillo, and elephant[1] [79]. This locus is representative of the whole genome, and contains coding, intergenic regions, and intronic regions. The multiple alignment of these regions, computed using TBA [17, 164], contains 1,000,000 columns. To simplify the calculations, consecutive alignment columns with the same gap structure were assumed to have undergone the same evolutionary scenario and were thus merged into a single "meta-column" we called an alignment *region*. Our alignment consisted of 123,917 such regions. Thus, during the execution of the Viterbi or Forward-Backward algorithm, the states are computed for each region instead of for each individual column, adapting the transition probabilities as a function of the width of each region. The phylogenetic tree used for the alignment and for the reconstruction is shown in Figure 7–1. The branch lengths are based on substitution rates estimated on a genome-wide basis [164]. For illustrative purposes, and similarly to the empirical values obtained by [130], the parameters of the indel model were set as follows: $\psi_D = 0.05, \psi_I = 0.05, \alpha_D = 0.9$, and $\alpha_I = 0.9$. However, we find that the ancestral reconstructions and confidence levels are quite robust with respect to these parameters (data not shown).

We first compared the maximum likelihood ancestral reconstruction found using our Viterbi algorithm to the ancestors inferred using the greedy algorithm of Blanchette *et al.* (2004) [16]. Table 3–2 shows the degree of agreement between the two reconstructed ancestors, for each ancestral node. We observe that both methods agree to a very large degree, with most ancestors yielding more than 99% agreement. The most disagreement concerns the ancestor at the root of the eutherian tree, which, in the absence of an outgroup, cannot

[1] In the case of cow, armadillo, and elephant, the sequence is incomplete and a small fraction of the bases are missing.

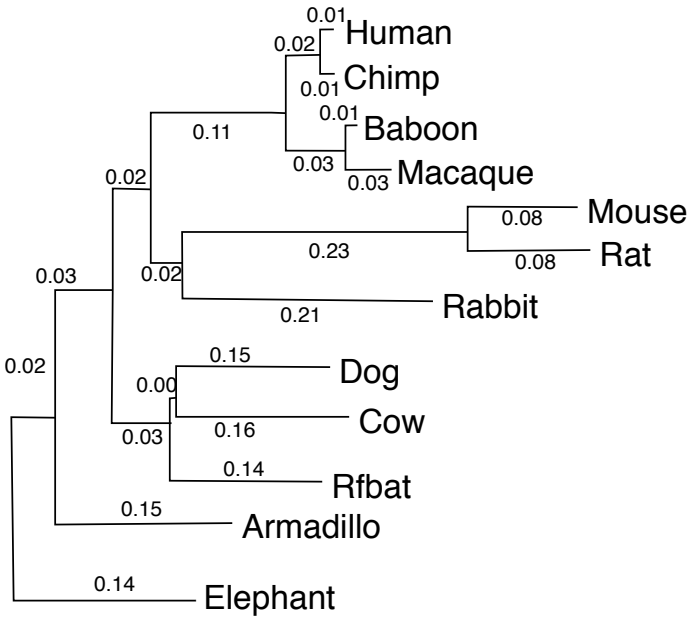


Figure 3–2: Phylogenetic tree for the twelve species studied in this paper.

be reliably predicted by any method. We expect that in most other cases of disagreement, the maximum likelihood reconstruction is the most likely to be correct, although the opposite may be true in case of gross model violations [116].

Table 3–2: Percentage of alignment columns where there is agreement between the ancestor reconstructed by the greedy algorithm of Blanchette *et al.* (2004) [16] and that predicted by our maximum-likelihood algorithm.

Ancestor	**% of agreement**
Mou+Rat	99.8181
Hum+Chi	99.9467
Bab+Mac	99.7275
Mou+Rat+Rab	99.8181
Hum+Chi+ Bab+Mac	99.7157
Hum+Chi+Bab+Mac+Mou+Rat+Rab	99.3901
Cow+Dog	99.917
Cow+Dog+Bat	99.8218
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat	99.0511
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat+Arm	93.6531
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat+Arm+Ele	84.9413

The main strength of the likelihood-based method is its ability to measure uncertainty, using the forward-backward algorithm, something that no previous method allowed. Assuming a phylogenetically correct alignment and a correct indel model, the probability that the maximum posterior probability reconstruction is correct is simply given by $\max\{\Pr(A_u^*[i] = 1|A), 1 - \Pr(A_u^*[i] = 1|A)\}$. For example, if $\Pr(A_u^*[i] = 1|A) = 0.3$, then the maximum posterior probability reconstruction would predict $A_u^*[i] = 0$, and would be right with probability 0.7. Figure 3–3 shows the distribution of this probability of correctness, for each ancestral node in the tree, over all regions of the alignment. We observe, for example, that 98% of the positions in the Boreoeutherian ancestor (the human+chimp+baboon+macaque+mouse+rat, cow+dog+rfbat ancestor, living approximately 75 million years ago), are reconstructed with a confidence level above 99% [2] . The ancestor that is the easiest to reconstruct confidently is obviously the human-chimp ancestor, where less than 0.14% of the regions have a confidence level below 99%. Again, the root of the tree is the node that is the most difficult to reconstruct confidently. Overall, this shows that most positions of most ancestral nodes can be reconstructed very accurately, and that we can identify the few positions where the reconstruction is uncertain. A potential drawback of the tree-HMM method is that its running time is, in the worst case, exponential in the number of sequences being compared. However, the optimizations described in this paper greatly reduce the number of states that need to be considered at each position, so the algorithm remains quite fast. Our optimized Viterbi algorithm produced

[2] We need to keep in mind, though, that these numbers assume the correctness of the multiple alignment, as well as that of the branch lengths and indel probability model, so that they do not reflect the true correctness of the reconstructed ancestor.

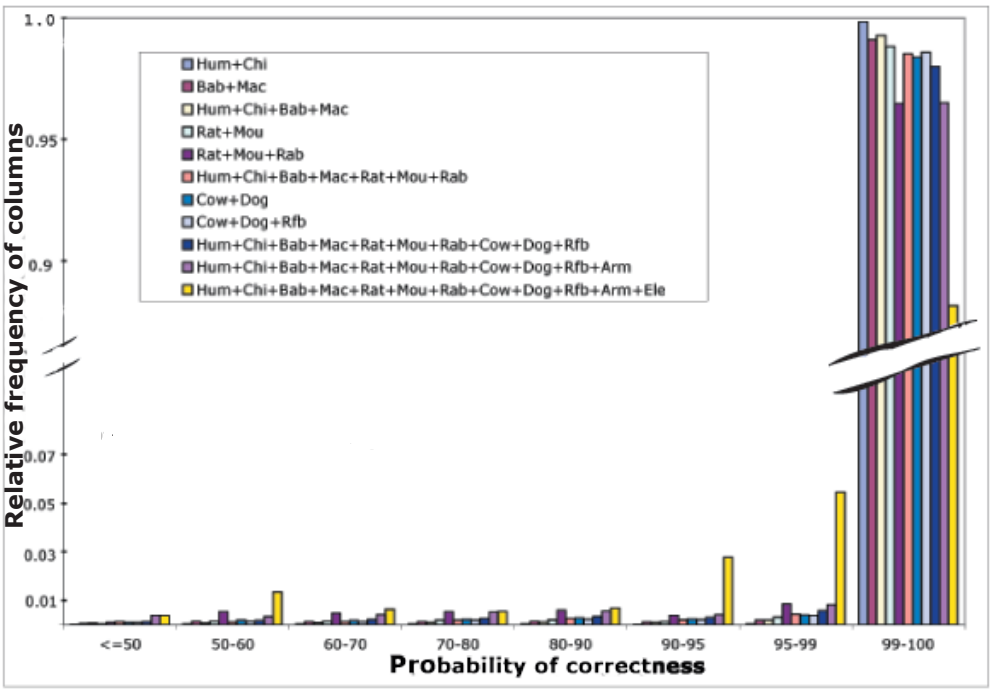


Figure 3–3: Distribution of the confidence levels, over all 123,917 alignment regions, for each ancestor. The vast majority of the ancestral positions are reconstructed with a probability of correctness above 99% (assuming the correctness of the alignment).

its maximum likelihood ancestral predictions on the 12-species, 1,000,000 column alignment in 7 hours on a Powerbook G5 machine, while the forward-backward algorithm produced an output after approximately double of that time. Figure 3–4 shows the distribution of the number of states that were actually considered, per alignment column. Most alignment columns are actually associated to less than 100 states. However, a small number of columns are associated to a very large number of states (15 regions have more than 100,000 states). Fortunately, these columns are rarely consecutive, so the incurred running time is not catastrophic for small number of species. However, to be applicable to complete genomes and to scale up to the more than 20

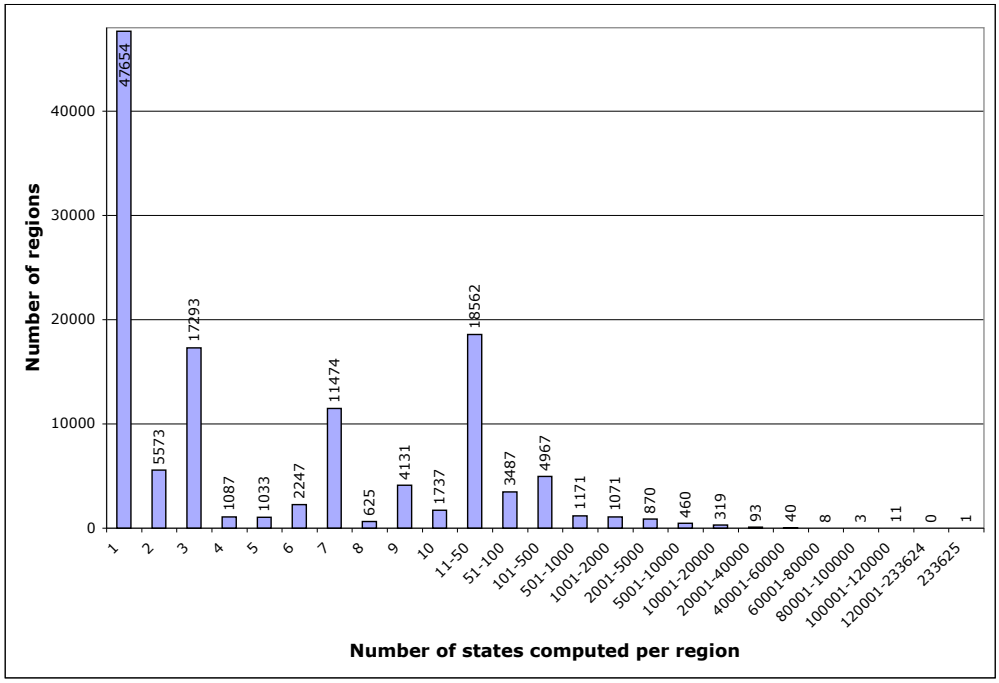


Figure 3–4: Distribution of the number of states considered ($|S'_i|$), over all 123,917 regions.

mammalian genomes that will soon be available, our algorithm requires further optimizations. These optimizations move away from an exact algorithm, toward approximation algorithms.

3.6 Heuristic algorithm for the IMLP

For each region i of the alignment and each possible state $s \in S'_i$, the exhaustive method considers all possible states for the next column, even though the Viterbi value $X(i, s)$ of some current state s may be far away from the maximal Viterbi value at that position, $\max_{s' \in S'_i} X(i, s')$. These states are less likely to be eventually chosen in the best path of the tree-HMM. Hence, to reduce the number of states created and reduce computation time, only states near the maximum Viterbi value are used to compute states for the next column. Thus, for region i, we distinguish between created states S'_i and used states $R_i \subseteq S'_i$, where only the second set will be involved in the creation of the states of the next column and in their Viterbi calculation. For position i,

state $s \in S'_i$ is retained in R_i if and only if $\log_2(\frac{\max_{s'} X(i,s')}{X(i,s)}) < t$, for some fixed threshold t. We note that this is equivalent to setting $X(i,s)$ to zero for each $s \in S - R$. A similar heuristic can easily be applied to the Forward-Backward algorithm. If t is sufficiently large, the loss in accuracy should be minimal for both algorithms, as will be shown next.

We computed the indels scenarios of the data sets presented in Section 3.5 by using different values for the threshold t. The approximate Viterbi algorithm was run using $t = 0, 1, 3, 5, 7, 9, 10, 20, 100$, and $+\infty$. Note that setting $t = 0$ results in a "greedy" algorithm that only considers the maximum Viterbi value at each position, while $t = +\infty$ give the original, optimal Viterbi algorithm. Figure 3–5 shows the number of states created (average of $|S'_i|$) and used (average of $|R_i|$) for all values of t, as well as the resulting running time. For small values of t, e.g. $t \leq 3$, only a handful of states are used, resulting in a very fast execution (less than 3 minutes). The average of number of states created increases relatively quickly with t, while the number of states used remains quite low (44.34 for $t = 100$). The average number of states created for $t = 20$ is about the same as the average number of states of the exact algorithm (see Figure 3–5), which shows that the used states are sufficient to give the necessary information to generate most valid states for next columns. Even though the average number of states created and used for $0 \leq t \leq 5$ is very low, the indels scenarios produced are very similar to the best scenario obtained by the exact method (see Table 3–3). We note that, for $t = 5$, the agreement with the exact algorithm is more than 99.99% for all the ancestors, while the running time is reduced by a factor of ten, and by a factor of one hundred for $t = 3$. For $t \geq 9$, the heuristic gives the optimal scenario, while still yielding a 5-fold speed-up. All values of t tested gave solutions that agreed with the optimal solution better than the solution

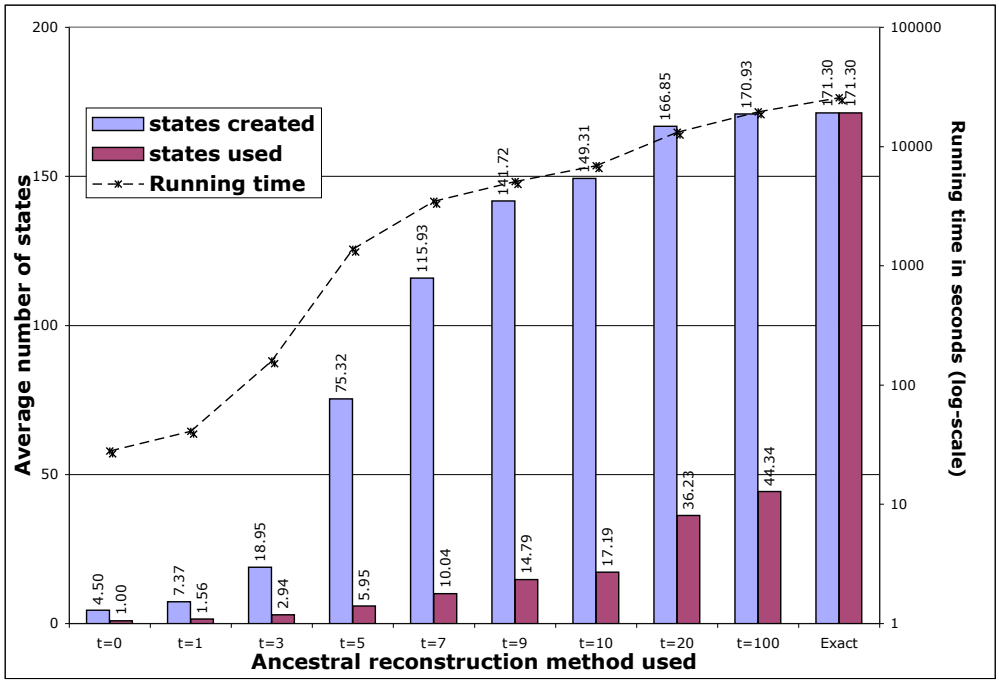


Figure 3–5: Average, over all alignment regions, of the number of states created (S'_i) and used (R_i), for the different values of the cutoff t. Running times (in seconds) are plotted with the log-scale shown on the right.

produced by the greedy algorithm of [16]. Finally, we note that, while our optimal Viterbi and Forward-Backward algorithms are limited to 12 to 15 species, our heuristic allows the inference of near-optimal solutions for much larger alignments. When run on a 1,000,000 column alignment of 28 species of vertebrates, our heuristic with $t = 3$ produced a solution in less than two hours. Since the exact algorithm cannot be run on such a large data set, it is difficult to estimate the quality of the solution obtained but, based on our experience on the smaller data set (Table 3–3), we expect a very high accuracy even at such a stringent cutoff.

Table 3–3: Percentage of alignment columns where there is disagreement between the ancestor reconstructed by the exact maximum-likelihood algorithm and the heuristic with different values for the cutoff t. We emphasize that the numbers quoted are percentages, so, for example, with $t = 0$, the Mouse+Rat ancestor agrees with the optimal solution at 99.97% of the alignment columns.

Ancestors	$t = 0$	$t = 1$	$t = 3$	$t = 5$	$t = 7$	$t > 9$
Mou+Rat	0.030	0.012	0.003	0.002	0.001	0
Hum+Chi	0.020	0.004	0.001	0.001	0.001	0
Bab+Mac	0.003	0.003	0.002	0.002	0.002	0
Mou+Rat+Rab	0.160	0.073	0.008	0.003	0.002	0
Hum+Chi+ Bab+Mac	0.060	0.041	0.011	0.002	0.002	0
Hum+Chi+Bab+Mac+Mou+Rat+Rab	0.160	0.070	0.018	0.006	0.004	0
Cow+Dog	0.070	0.032	0.006	0.002	0.001	0
Cow+Dog+Bat	0.080	0.049	0.013	0.002	0.001	0
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat	0.170	0.095	0.017	0.005	0.004	0
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat+Arm	0.100	0.048	0.010	0.003	0.002	0
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat+Arm+Ele	0.010	0.004	0	0	0	0

3.7 Discussion and Future Work

The method developed here allows predicting maximum likelihood indel scenarios and their resulting ancestral sequences for large alignments. Furthermore, it allows the estimation of the probability of error in any part of the prediction, using the forward-backward algorithm. Integrated into the pipeline for whole-genome ancestral reconstruction, it will improve the quality of the predictions and allow richer analyses. The main weakness of our approach is that it assumes that a phylogenetically correct alignment and an accurate phylogenetic tree are given as input. While many existing multiple alignment programs have been shown to be quite accurate on mammalian genomic sequences (including non-functional or repetitive regions) [17], it has also been shown that a sizeable fraction of reconstruction errors is due to incorrect alignments [16]. Ideally, one would include the optimization of the alignment directly in the indel reconstruction problem, as originally suggested by Hein in 1989 [107]. However, with the exception of statistical alignment approaches [152], which remain too slow to be applicable on a genome-wide scale, genomic multiple alignment methods do not treat indels in a probabilistic framework. We are thus investigating the possibility of using the method proposed here to detect certain types of small-scale alignment errors, and to suggest corrections.

When predicting ancestral genomic sequences, it is very important to be able to quantify the uncertainty with respect to certain aspects of the reconstruction. Our forward-backward algorithm calculates this probability of error for each position of each ancestral species. However, errors in adjacent columns are not independent: if position i is incorrectly reconstructed, it is very likely that position $i + 1$ will be wrong too. We are currently working on models to represent this type of correlated uncertainties. This new type of

representation will play an important role in the analysis and visualization of ancestral reconstructions.

Finally, it will be important to assess the results given by the heuristic so that the cutoff value t is chosen appropriately for the data at hand. For example, the heuristic could be applied iteratively by increasing the cutoff until a stationary likelihood score is reached. This heuristic will be useful to reconstruct the indel scenarios for data sets containing more than 20 taxa and could be easily applied to the large number of mammalian genomes that are about to be completely sequenced.

3.8 Acknowledgements

A.B.D. is an NSERC fellow. We thank Éric Gaul, Eric Blais, Adam Siepel, and the group of participants to the First Barbados Workshop on Paleogenomics for their useful comments. We thank Webb Miller and David Haussler for providing us with the sequence alignment data.

CHAPTER 4
Web Tools for Indel and Ancestral Sequence Reconstruction

In this chapter, we present two web tools intended to contribute to the ancestral sequence reconstruction project. The first allows the visualization of the n best indel scenarios for a given set of aligned sequences [56]. The second is dedicated to a web interface allowing one to compute three important steps of the ancestral genome reconstruction project (multiple sequence alignment, indel inference and substitution inference) [59].

4.1 Visualizing the n best Indels scenario in Ancestral Genome Reconstruction

4.1.1 Preface

This section contains a part of the following paper: **Diallo, A.B.**, Gaul, E. (2009): Visualization of the n best indel likelihood scenarios. *In preparation.* 14 pages.

4.1.2 Abstract

The reconstruction of indel evolutionary scenarios often leads to several solutions. These solutions may disagree for several blocks of characters in the reconstructed ancestors. Although two solutions may be equally parsimonious, the likelihood associated to the reconstructed blocks where the divergence occurs can be different, depending on where one looks in the tree. Diallo et al. 2007, in addition to describing a method to compute the most likely indels evolutionary scenarios, showed that the likelihood of each ancestral character (position) can be assessed to be present or not [58]. In this paper, we propose an efficient way of visualizing ancestral reconstructions and the likelihood associated to each of the reconstructed blocks. We establish a way

of presenting the conflicting solutions in a single figure while highlighting the regions that have a high/low likelihood. The tool we propose displays the set of conflicting solutions using a graph to represent each ancestral sequence or each operation that can explain transitions between two nodes. The tool is web-based, and the navigation within the figures is intuitive.

4.1.3 Introduction

Given a phylogenetically-correct multiple sequence alignment, there are two main approaches to infer evolutionary scenarios of insertions and deletions: the most parsimonious scenario and the most likely indel scenarios. The reconstruction of the most parsimonious scenario has been studied by [35, 91]. Recently, two algorithms have been described for the maximum likelihood reconstruction ([62, 131]), which is preferable to get a broader view of the process, but also presents significant algorithmic challenges. In this paper, we proposed an extension of [62], allowing one to compute and to visualize the results of the most likely indel scenarios. While several indel solutions could have the same parsimony score or likelihood score, part of their solution might differ. Hence, the likelihood associated to each part of a given solution could be a *good* measure for the comparison between competing solutions. In [62], a tree-based hidden Markov model [217] was used to identify the most likely indel scenario. Here, the standard Viterbi and Forward-backward algorithms used to produce the most likely ancestral reconstruction has been extended to allow the computation of the $n-$best indel scenarios and to compute the likelihood associated to specific regions of the solution. We also define an efficient way to reconcile those solutions, or highlight possible disagreements between them. For this purpose, the following probabilities and likelihoods need to be computed: the probability of either having a nucleotide or not at each position of the ancestor, the likelihood of the transition from a character

(gap or nucleotide) into another character at neighboring position. The proposed tool presents the n best indel solutions for the entire set of all ancestral nodes. Highlights of the consensus in each part of the competing solutions have been presented. Other informations, such as the sequence of operations between an ancestors and its descendants, are also presented. The designed tool is web-based, and uses a simple navigation scheme to avoid overloading the user with information. Here, we used intuitive concepts from information theory to simplify the interpretation of the results. The proposed tool allows users to grasp at a glance the likelihood of the best solution, and to decide if the potential conflicts between other possible scenarios and the best one are worth considering or not.

4.1.4 Computing the n best paths using the Viterbi algorithm

The standard Viterbi algorithm (see [67]) computes the best sequence of states that explains an observed sequence of emitted character from a given hidden Markov model. One can adapt the algorithm, while increasing temporal and spatial complexity, to obtain the n best paths, where n is a parameter specified by the user. More precisely, the standard Viterbi algorithm maximizes the value $P(\pi|M)$, where M is the given sequence of observation and π is a sequence of states that can emit M. It works by filling the ***Viterbi table*** of ***Viterbi values*** $v_k(i)$ that is indexed by the states of the hidden Markov model (e.g. k) and the given positions (columns, e.g. i) of the emitted sequence. The value at the last column represents the maximal value for $P(\pi|M)$. Then, the most likely sequence of states can be retrieved by a trace-back approach. However, to keep the $n-$best paths, we must also be aware that each pair state-column may be part of one or many of these best paths. The intuitive approach is to keep track of the n best Viterbi values at each

state-column:

$$v_k^1(i-1), v_k^2(i-1), \ldots, v_k^n(i-1).$$

A back pointer to the state-column-order of the last column that led to the current value can be recorded to facilitate the traceback of the paths (see Figure 4–1). Hence, each Viterbi value can be computed according to the following formula:

$$v_k^\alpha(i) = e_k(i) \times \alpha^{th} bestscore_{l,h} v_l^h(i-1) a_{lk}, \tag{4.1}$$

where $e_k(i)$ is the emission probability of column i from state k, and a_{lk} is the transition probability from state l to state k, both values being given by the model.

The presented approach is expensive due to a n-fold increase of both space and time. However, several optimizations could take place such as such as those presented in [37].

Having the n best paths, we can now extract from each path the reconstructed sequence for a given node or branch, and merge the information into

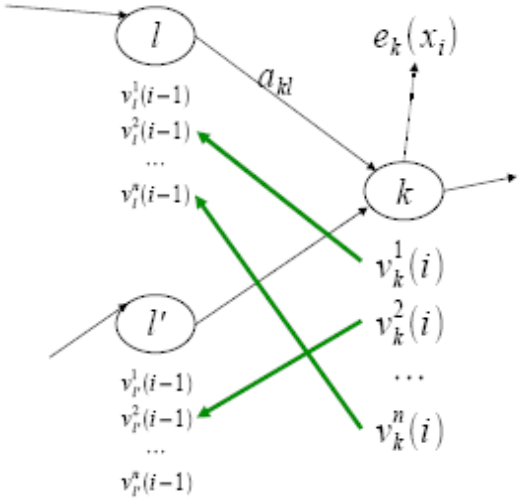


Figure 4–1: Computation of the n best paths with Viterbi algorithm.

a *consensus Directed Acyclic Graph* to summarize consensus or disagreement among optimal solutions for the corresponding node or branch.

Definition 4.1 (consensus Directed Acyclic Graph) *The consensus directed acyclic graph is a graph representing several conflicting indel scenarios for a given node or branch. Given n indel scenarios from a multiple sequence alignment A and phylogenetic tree T, a directed acyclic graph G_b representing the possible subset of $\{0,1\}^L$ sequences given by the indel scenarios for a node b of T is computed as follows:*

- *Create* begin *node and* end *node to represent the character position of respectively the first column and the last column of the alignment.*
- *Build a single node for each consecutive characters shared by a set of solutions for a given region of the alignment.*
- *Build individual nodes for a given region of the alignment with different characters in the set of solutions.*
- *Build arcs between nodes such that each indel solution is represented by a path from the* begin *node and the* end.
- *For each node, the corresponding sequence of characters is used as label.*
- *Each label of an arc contains the set of identification of the indel solutions that have a path traversing this arc.*

One can notice that merging the labels of the nodes for a given path of a directed acyclic graph gives an ancestral sequence for the corresponding indel scenarios. Similar directed acyclic graphs can also be built for the branches of the phylogenetic tree according to the indel evolutionary scenarios on branches. An example of a consensus directed acyclic graph is presented in Figure 4–3.

It is important to notice that, for a given set of best indel scenarios, most of the best paths could be share large regions on a given node. The differences between the ancestral sequences given by the different solutions are mostly

reflected by a single character difference, in different positions. Moreover, representing these kind of solution for large genomic sequences will lead to directed acyclic graphs that are difficult to visualize and to interpret. Hence, in this approach, we required the divergences between solutions to be present in several consecutive columns to be taken into account. The threshold used in this study is seven consecutive characters.

4.1.5 Assessing uncertainties

The goal of the visualization tool is to highlight the regions for which it is more difficult to establish a consensus solution. Here we present how the uncertainty can be computed and represented in the obtained indel results. For this purpose, we separate this task into two parts: how to compute the uncertainty from the model and represent it in the consensus directed acyclic graph (for detailed presentation), and how to represent this uncertainty for reconstructed sequences on the global tree (for fast observation).

Computing the reconstruction uncertainties and representing it in the consensus Directed Acyclic Graph. From a consensus directed acyclic graph, we would like to answer the following two questions. Given an evolutionary model of indel, a node (or the branch) of the tree, the alignment of extant species and a position in the given alignment:

- What is the probability to have a character at the given position in the given node of the reconstructed sequence ?
- What is the probability of having a nucleotide at the next position given that the current position of the ancestor sequence has a nucleotide?

The answer to the first question can be obtained by computing for each ancestral node A_l, the likelihood of having a nucleotide at position i, defined as

follow:

$$P[A_l(i) = 1|Alignment] = \sum_{\{k|k \ predicts \ 1 \ at \ A_l(i)\}} P[\pi_i = k|Alignment] \quad (4.2)$$

Application of the *forward-backward* algorithm [67] gives the answer (where the paths for which we compute the forward and backward values are restricted to those that pass through k on column i). To compute the probabilities of transitions, the following normalization scheme, computed from the Bayes formula, is suitable:

$$P[A_l(i)|A_l(i-1), Alignment] = \frac{P[A_l(i-1), A_l(i)|Alignment]}{P[A_l(i-1)|Alignment]} \quad (4.3)$$

Representing uncertainties of reconstructed sequences in the global tree. The events that happen on edges of the input tree, at a certain position of the alignment are represented as four operations (conservation, insertion, deletion, and no event), showed using characters C, I, D and $*$. A straightforward way to see rapidly if the best solutions agree together is to compute what we call the agreement of the events for a branch at a certain position. Given the probability distribution p_C, p_D, p_I, p_* of the characters at this point (over all paths of the HMM), the agreeement among them is defined as follows:

$$Agreement = 1 - Normalized \ entropy = 1 - \frac{-\sum_{\{i \in \{C,D,I,*\}\}} p_i log(p_i)}{log(4)} \quad (4.4)$$

The intuitive interpretation of the agreement is that it will be the highest (i.e. 1) if we are sure that the identified character is the right one, and the lowest (i.e. 0) if the uncertainty is the highest. To represent the agreement for all positions of a sequence of events, we use a histogram, that is attached to the corresponding branch of the phylogenetic tree. The user can instantly

appreciate the degree of discrepancy in the reconstructed regions, by identifying zones in the histograms where the agreement between retrieved operations given by the indel scenarios is low. For the characters, the probability to have a 1 at a certain position in the given ancestor (node) is represented using a level of shadding.

4.1.6 Visualizing the reconstruction and the uncertainties

Technical considerations. The tool was developed in Java for web purposes. Each figure corresponds to a generated PNG file, with an associated map that defines the boundaries of the clickable elements within the figure. The GraphViz package has been used to generate the graphs, and JFreeChart to generate histograms. An output sample is available at: <http://www.mcb.mcgill.ca/~egaul/viagr/>.

Interpretation of the program output. Figures 4–2 and 4–3 presents the visualization of the results obtained from the reconstruction of the 40 best indels scenarios of eight mammalian sequences and 1kb alignment size. A quick glance at the red histograms directly under the root (node 7 of Figure 4–2) tells us that the problem is difficult for the first third of the alignment, which is not surprising, given the little information contained in this area of the alignment (see Figure 4–4). However, there is a consensus for the reconstructions of the last two-third of the sequences (the ancestor seems to be mostly composed of characters in that area).

Clicking on node 7 in the tree of Figure 4–2 gives a detailed view as presented in Figure 4–3. Hence, one can visualize the details of the competing paths represented here. The values computed by the previous formulas are available by pointing any nodes, as well as any edges of the consensus directed acyclic graph. Since determining the n best paths is a pre-processing step

to filter out uninteresting reconstructions, one should notice that not all the values are presented, but only the most interesting ones.

4.1.7 Conclusion

We have proposed a web-based tool to visualize the results of the reconstruction of most likely indels scenarios, given an alignment of extent sequences. One of the major assets of this tool is to present uncertainty on the reconstruction in a clear and intuitive way, by histograms presenting agreement between most likely scenarios, and consensus directed acyclic graphs presenting consistent and/or conflicting zones in reconstructed sequences. Another important characteristic of the tool is to facilitate visualization of the results

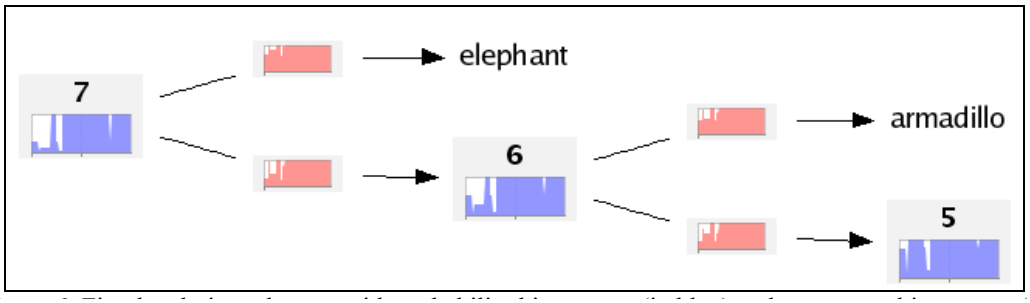


Figure 4–2: First level view: part of the tree with probability histograms (in blue) and agreement histograms (in red). We see at a glance that the example problem is hard for the first third part of the alignment. One has to click on the histogram to visualize the consensus directed acyclic graph (e.g. node 7 is represented in Figure 4–3).

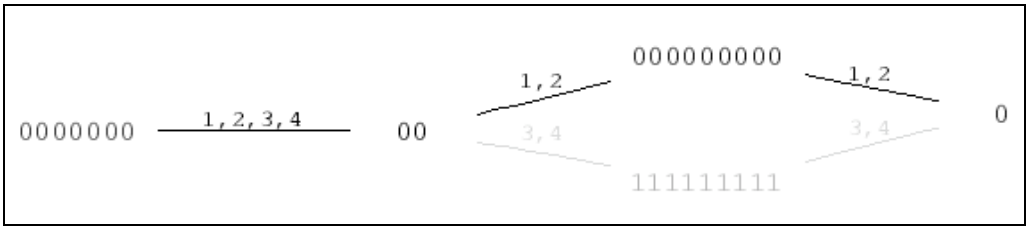


Figure 4–3: Consensus Directed Acyclic Graph summarizing four paths from the best ones in the reconstruction of the root node (7) (view of columns 2 to 20 from the example presented in the result). The number on the edges represent the path identifications.

```
human        A--------- ATTGTAACCA TTCATT---- --TAATCTGA CTACTTCCCT
chimp        ANNNNNNNNN ATTGTAACCA TTCATT---- --TAATCTGA CTACTTCCCT
mouse        A--------- ---------A TTTATTTA-- --TAATTTGA CTGCTTATAT
rat          ATACTAGC-- ---------A TTTATTTA-- --TAATTTGA CTGCTTACAT
dog          A-------AT ATTGCAATTA TTTATT--TA TACAGTCTAC CTTCTT-CAC
cow          ANNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN
armadillo    ANNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN
elephant     ANNNNNNNNN NNNNNNNNNN TTCATA---- --GAATCTGA CTACTTCCAC

             TGGATCATTA CTGACACTAG TGAACAACT- CTTTTTCATC TCCTTTGTA
             TGGATCATTA CTGACACTAG TGAACAAAT- CTTTTTCATC TCCTTTGTA
             TGGATCATTA TTGAAACCAG TGATTAAATT TTTTTTAATA TCTCATGAA
             TG-ATCATTA CTGAAACCAG TGAGTAAAT- ---------- ---------
             TGGATCATGA CTGAAACTAG GGAACAAAT- CCTTTTCCTC TCCAATGTG
             NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNN
             NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNNN NNNNNNNNN
             TGGATCATTA CTGAAACTAG GAAACAAGT- CTTTTTCATC TCCAGTGTG
```

Figure 4–4: Part of the sequence alignment of eight mammals given as input of the indels scenarios for the visualization of the Figure 4–2.

themselves by displaying them along the phylogenetic tree in a graphical-oriented way. Some problems are still to be resolved, the most striking one being the long processing time between the moment the user inputs an alignment and the moment when results are displayed. However, the tool is still usable in a batch-oriented system. It is also important to notice that we presented the results according to the n best paths. However, it would be possible, instead of summarizing the n best paths in the consensus directed acyclic graph, to create one that would keep only nodes and/or edges for which the probabilities given by the result of the *forward-backward* algorithm are above a fixed threshold. We plan also to extend this tool to offer a way to zoom and choose the detail level in which the user wants to navigate. Hence, it would be possible to track large zones for which there is a reconstruction conflict, as well as localize the zone with low level of confidence.

4.2 Ancestors 1.0: A web server for the ancestral genome reconstruction

4.2.1 Preface

This section contains the following paper ready to be submitted : **Diallo, A.B.**, Makarenkov, V., Blanchette, M. (2009): Ancestors 1.0: A web interface for ancestral sequence reconstruction. Bioinformatics, Application Note. *In preparation.*

4.2.2 Abstract

Ancestral genome reconstruction is composed of five difficult steps: Identifying syntenic regions, inferring ancestral arrangement of syntenic regions, aligning orthologous sequences, reconstructing insertion and deletion histories, and finally inferring substitution histories. Here, we present a web server allowing one to easily and quickly compute the last three steps of the ancestral genome reconstruction procedure. We implemented several alignment methods, an indel maximum likelihood solver and the context-dependent Felsenstein algorithm for predicting substitutions. The results presented by the server include the posterior probabilities for the last two steps of the ancestral genome reconstruction and the error rate of each ancestral base prediction. This server is available at the following URL address:
<http://ancestors.bioinfo.uqam.ca/ancestorWeb/>.

4.2.3 Introduction

Comparative genomics is a field that uses the information provided by the patterns of selection to understand the functions and the evolution on different genomic regions. Studies in comparative genomics are often based on a direct analysis of multiple sequence alignments of extant sequences [208]. However, the recent interest for ancestral genome reconstruction also provides clues on

several aspect of evolutionary changes [15, 16]. Ancestral genome reconstruction attempts to predict the DNA sequences of all ancestral species in a given phylogeny according to a multiple sequence alignment. Accurate ancestral genome reconstruction can contribute to the study of adaption, behavioral changes and functional divergence [15, 16, 58]. Two of the most important steps in ancestral genome reconstruction procedure are the prediction of substitution and insertion and deletion (indel) that may have produced a given set of aligned regions [16]. Although the inference of indel evolutionary scenarios is useful in several problems, it has received relatively little attention. We have recently proposed a statistical framework that enables one to infer the most likely indel scenario and to estimate uncertainties of predictions based on fixed indel rate parameters and a given multiple sequence alignment. The developed framework is adequate for small-scale genomic regions with insertions, deletions and substitutions [58, 62]. Substitutions are predicted using an adaptation of the Felsenstein algorithm [82, 88] described in [16]. The maximum likelihood indel scenario is predicted by an exact algorithm described in [62]. This algorithm uses a special type of hidden Markov model [194], called tree-HMM, which is a combination of a standard hidden Markov model and phylogenetic trees [88, 257]. We showed that the most likely path through the tree-HMM leads to the most likely indel scenario and that a variant of the standard Viterbi algorithm [67] can solve the problem. The uncertainty associated to each of the reconstructed ancestral regions can be assessed using the standard Forward-Backward algorithm [67].

The Ancestors web server presented in this paper performs the last three steps of the ancestral genome reconstruction procedure. It allows to compute multiple sequence alignments using several widely used algorithms, to infer exact or heuristic based indel scenario and to predict substitutions. All the

steps have been combined in a single web interface. The results are presented as colored output indicating the level of confidence of each prediction. Ancestors 1.0 is available at the following URL address:
<http://ancestors.bioinfo.uqam.ca/ancestorWeb/>.

4.2.4 User inputs and Ancestors 1.0 outputs

The Ancestors 1.0 interface can be divided in two parts. The first concerns the user inputs to the program through a web form. The second one is dedicated to presenting the obtained results. At each step, the user can send via an e-mail questions or feedback to the system administrator. Sample sets are also ready to test. Moreover, there are links to the required formats for the different inputs and an explanation of the different sections of the results.

User inputs. The Ancestors 1.0 web form is divided in three parts (see Figure 4–5). The first part allows users to supply a set of orthologous sequences in Fasta format. The sequences could either be aligned or not. Users can choose a method of sequence alignment even though the sequences are already aligned. Realigning sequences could help increasing the alignment accuracy. The following multiple sequence alignment procedures are available: Clustal-W [112], Dialign [170], Mavid [25] and TBA [17]. Those methods have been chosen for one or more of the following reason(s): they are widely used in comparative genomics, and they have been shown to be accurate on genomic data, they can handle a reasonable large datasets.

The second part allows users to supply a rooted phylogenetic tree related to the multiple sequence alignment data that can be used to guide the ancestral sequence reconstruction. The tree format is the widely used Newick format. The third part concerns the choice of the indel approach to use and the related parameters. There are parameters related to the tree-HMM and other parameters related to either the exact algorithm or the heuristic one.

Users can either ask to report the most likely indel scenario or the posterior decoding for predicting the presence or absence of characters at each position of the ancestral sequences. The posterior decoding can be used to compute the confidence levels of the predictions as described in [58]. The substitution parameters are those of the HKY [105] evolutionary model. Future versions of the program will allow the reconstruction of the phylogenetic tree by supplying several phylogenetic tree reconstruction methods. Moreover, the form should allow unrooted tree as input and proposed ways to rooted it.

Ancestors 1.0 outputs. The Ancestors 1.0 results are made available using different plain text and HTML format files. The summary of those files is given as default output (see Figure 4–6). The results present each input file, followed by the result of the alignment method (if an alignment procedure has been chosen). Then, the summary of the different command lines executed is given as a plain text. The results of the indel predictions are presented in three files (the indel scenario, the tree-HMM state created, the posterior probability for each position of each sequence).

The results of the ancestral nucleotides predicted contain the characters in plain text and HTML file. The HTML file presents a colored output according to the level of prediction confidence, as shown in Figure 4–7. Those confidence levels are given in separate file together with the obtained posterior probabilities.

4.2.5 Ancestor Help

Users can consult the *Ancestors 1.0 user guide* available at the web site for detailed explanation of the file formats present in Ancestors 1.0. For specific help request, users can use the bug report form.

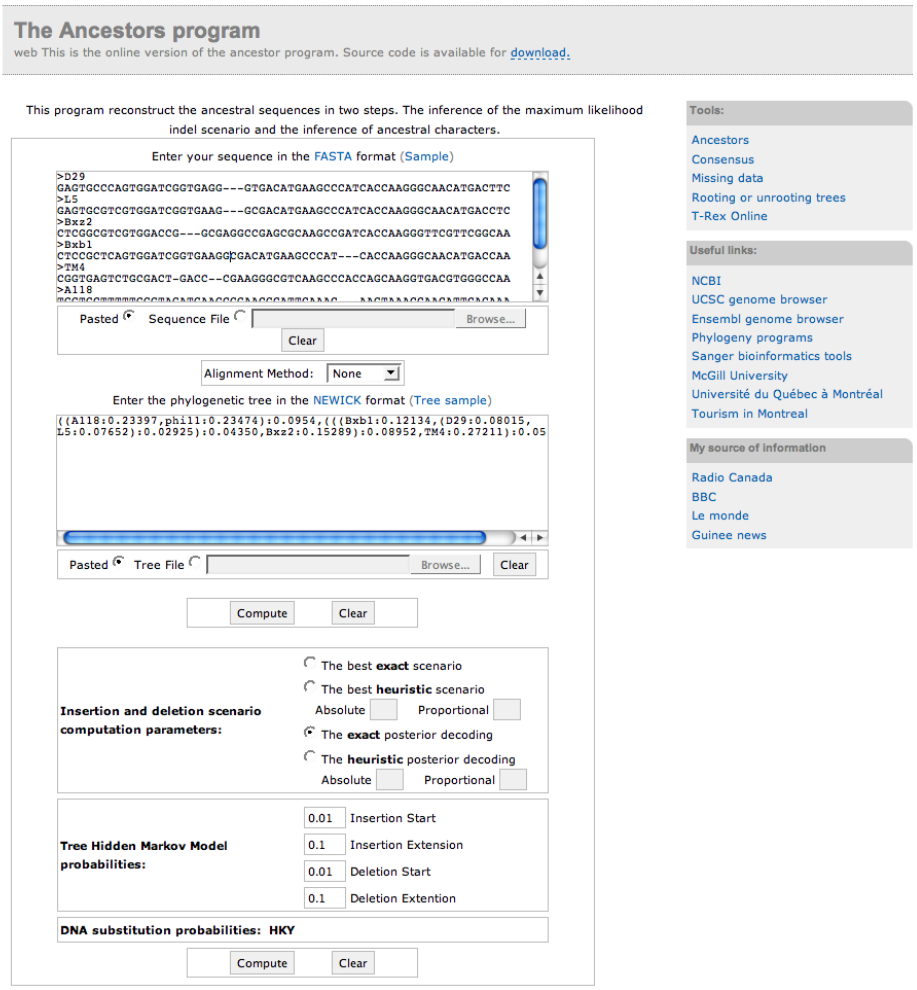


Figure 4–5: The Ancestors 1.0 user input form. The interface consists of three parts: the alignment, the phylogenetic tree and the indel parameters. On the right of the display, links are given to commonly used tools and databases in comparative genomics. The alignment and the tree present a set of phage genes.

4.2.6 Acknowledgement

We thank Alix Boc and Nathalia Vilgrain for their help in developing the web interface and the referees who will test the web server. A.B. Diallo is a NSERC fellow.

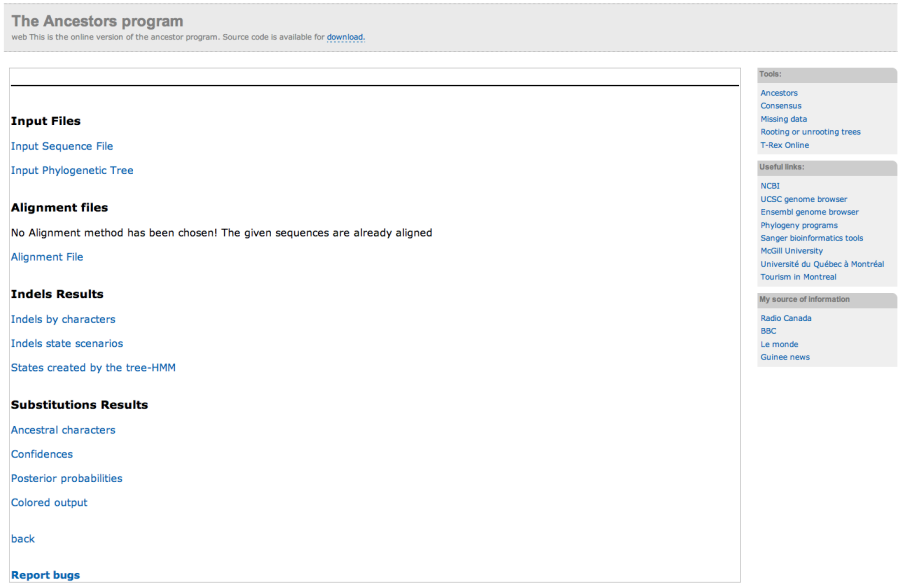


Figure 4–6: The Ancestors 1.0 main output. It gives the links to all the obtained result files as well as input files.

```
> A118+PHI11+
TGGTGCTTATTGGGTAGGTGAAGGGGATCACATCAAGACCACAACTAAGGCTACATTGGCCAA
> D29+L5+
GAGTGCTTAGTGGATCGGTGAAG---GCGACATGAAGCCCATCACCAAGGGCAACATGACCTC
> BXB1+D29+L5+
CAGTGCTTAGTGGATCGGTGAAGGGGACGACATGAAGCCCATCACCAAGGGCAACATGACCAA
> BXB1+D29+L5+BXZ2+
CAGTGCTTAGTGGATCGGTGAAGGGGACGACATCAAGCCCATCACCAAGGGCACCTTGGCCAA
> BXB1+D29+L5+BXZ2+TM4+
CGGTGCTTATTGGATCGGTGAAGGGGACGACATCAAGCCCACCACCAAGGCTACCTTGGCCAA
> A118+PHI11+BXB1+D29+L5+BXZ2+TM4+
CGGTGCTTATTGGATAGGTGAAGGGGACGACATCAAGCCCACCACCAAGGCTACCTTGGCCAA

Confidence levels: 0                                    100
```

Figure 4–7: The ancestral sequence predictions and the corresponding confidence level (between 0 and 100) of each character. These confidence levels have been computed according to the confidence level of the indel predictions as well as the substitution predictions. The ancestral names correspond to a concatenation of the children names.

CHAPTER 5
Evolutionary Score for the Multiple Sequence Alignment refinement and the Joint Inference of Phylogenies and Multiple Sequence Alignment

5.1 Preface

In this chapter, we show a way of using the previous framework for indel maximum likelihood problem to compute an indel likelihood score that can be used for phylogenetic tree inference and alignment quality assessment. This score is valuable for refining alignment and trees as well as doing the joint inference of both. The text presented here is taken from Diallo et al. [60], (*in preparation*).

5.2 Abstract

Recently, several papers pointed out the necessity of the joint inference of phylogenetic trees and multiple sequence alignments. Such an inference will tend to avoid the cycle of direct dependency of both reconstructions [71, 72, 107]. However, a joint inference requires the availability of an adequate criterion of assessing the agreement between a phylogenetic tree and a multiple sequence alignment. Here, we propose to use the indel likelihood score as a measure of such an agreement. This indel likelihood score can be computed efficiently using existing tree-HMM approach proposed by Diallo and coauthors in 2007 [58]. Monte Carlo simulations were realized to obtain synthetic multiple sequence alignments and phylogenetic trees. The simulation results showed that the indel likelihood score correlates well with the accuracies of both phylogenetic tree and multiple sequence alignment.

5.3 Introduction

Comparative genomics is a field that uses the information provided by the patterns of selection to understand the functions and the evolutionary processes on different genomic regions. Studies in comparative genomics (as well as several other areas in computational biology) are often based on a direct analysis of multiple sequence alignments [208]. However, the recent interest for ancestral genome reconstruction also contributes to get clues on several aspects of evolutionary changes [15]. Ancestral genome reconstruction attempts to predict the DNA sequences of all ancestral species in a given phylogeny according to a multiple sequence alignment. One of the most important steps in the ancestral genome reconstruction procedure is the prediction of the set of substitutions, insertions and deletions that may have produced a given set of aligned regions [16]. Although the inference of insertion and deletion (indel) evolutionary scenarios is useful in several problems, such as annotating functional genomic regions [208, 215], it has received relatively little attention. We have recently proposed a probabilistic framework that allows to infer the most likely indel scenario and estimate uncertainties of predictions based on a given phylogenetic tree and fixed indel rate parameters. The developed framework is adequate for small-scale genomic regions with insertions, deletions and substitutions [58]. It makes the assumption of having correct multiple sequence alignment and phylogenetic tree as input. While many existing multiple sequence alignment and phylogenetic tree programs have been shown to be quite accurate on different genomic sequences [25, 72, 129], it has also been shown that a large fraction of ancestral sequence reconstruction errors is due to incorrect alignments [15]. The problem of having a correct alignment and an accurate phylogeny is a significant issue in computational biology [72]. Most of the existing approaches for building multiple sequence

alignment and a phylogenetic tree from a given set of sequences first build a multiple sequence alignment and then reconstruct the phylogenetic tree based on this multiple sequence alignment. However, such a direct dependency of multiple sequence alignment and phylogenetic reconstruction can lead to biased estimations. Thus, the ideal solution is the joint inference of both of them. Ideally, one would include the optimization of the alignment and phylogenetic tree directly in the indel reconstruction problem, as originally suggested by Hein [107]. However, with the exception of statistical alignment approaches [107], which remain too slow to be applicable on a genome-wide scale, few genomic multiple sequence alignment methods treat indels in a probabilistic framework. Existing methods for the identification of indel and substitution scenarios make the assumption of having as input a correct alignment and an accurate phylogeny. However, several studies try to overcome this issue by doing successive alternate refinements of the multiple sequence alignment and the phylogenetic tree. The multiple sequence alignment and the phylogenetic tree are corrected in separate frameworks. Seminal contribution makes an iterative refinement of the tree and the multiple sequence alignment [72] or only a refinement of the multiple sequence alignment [131]. The difficulty of handling both in the same algorithmic schema is due in part to unavailability of a quick estimator of the joint accuracy of the multiple sequence alignment and the phylogenetic tree. In this paper, we propose to use the evolutionary likelihood score as an estimate of both correctness. Moreover, the evolutionary score can be used to choose the best phylogenetic tree or the best alignment. Previous studies on defining likelihood scores often rely only on substitutions, except for an initial contribution implemented in Indelign program[131]. However, the computational requirement of the latter program makes it unpractical for large datasets such as analyzing long regions of several sequenced mammalian

genomes. In this paper, we showed how this score can be computed easily using the ancestors framework [58]. Moreover, we demonstrate, using simulation studies, how it agrees and fits well with multiple sequence alignment correctness and phylogenetic tree accuracy.

5.4 Indel Likelihood Score

In this section we will give a precise definition of the indel likelihood score. Consider a rooted binary phylogenetic tree $T = (V_T, E_T)$ with branch lengths $\lambda : V_T \rightarrow \mathbb{R}^+$. If n is the number of leaves of T, there are $n-1$ internal nodes and $2n-2$ edges. Consider a multiple alignment A of n orthologous sequences of size L corresponding to the leaves of the tree T. An ancestral reconstruction corresponds to an extension of A that assigns a sequence $A^*_u \in \{1,0\}^L$ to each node u of T, and where $A^*_u = A_u$ whenever u is a leaf. Notice that a pair of binary alignment rows A^*_x and A^*_y can be partitioned into a set of non-overlapping insertions, deletions, and conservations. Following the rules of phylogenetic correctness defined in [58], let Γ be the set of all correct reconstructions of A^*.

Consider an evolutionary model describing the probabilities of indels along each edge of the phylogenetic tree. This defines the probability of transition between an alignment row A^*_x and its descendant row A^*_y. This probability is a function of the probability of the insertions, deletions, and conservation that happened from A^*_x to A^*_y. It can be defined as follows:

$$\begin{aligned}\Pr(A^*_y|A^*_x, b) \;=\; & \prod_{(s,e)\text{: deletion from } A^*_x \text{ to } A^*_y} P_{Deletion}(s,e,b) \cdot \\ & \prod_{(s,e)\text{: insertion from } A^*_x \text{ to } A^*_y} P_{Insertion}(s,e,b) \cdot \\ & \prod_{(s,e)\text{: conservation from } A^*_x \text{ to } A^*_y} P_{Conservation}(s,e,b)\end{aligned} \tag{5.1}$$

where (s, e) delimits the position start s and end e of a specific region; $P_{Insertion}(s, e, b)$ and $P_{Deletion}(s, e, b)$ represent the probability of respectively inserting and deleting a region of size $e-s+1$ according to the branch length b; $P_{conservation}(s, e, b)$ corresponds to the probability of the characters to be conserved according to the branch length b. Hence the likelihood of an ancestral reconstruction A^* for A can be formulated as :

$$L(A^*) = \prod_{b=(x,y)\in E_T} \Pr(A^*_y | A^*_x, b). \tag{5.2}$$

Given T, A, and an evolutionary model Θ, the indel likelihood score can be defined as:

$$Is(T, A, \Theta) = \sum_{A^*\in\Gamma} L(A^*) \tag{5.3}$$

In a previous publication, we solved the indel maximum likelihood problem using a tree-HMM [58]. This tree-HMM represents all the possible phylogenetic correct indel scenarios, where each correct indel scenario refers to an unique assignment $A^* \in \Gamma$. Hence, the execution of the standard forward algorithm in the tree-HMM will sum all the possible assignments $A^* \in \Gamma$ as required to compute $Is(T, A, \Theta)$ [67].

5.5 Simulation Procedure

A Monte Carlo study was conducted to test the ability of our indel score to measure the agreement between a phylogenetic tree and a multiple sequence alignment. For this purpose, the mammalian tree given in Figure 5–1 was used to simulate the evolution of a known (but synthetic) ancestral sequence through the different lineage of the phylogenetic tree, using the simulation program Simali (*http://www.bx.psu.edu/miller_lab/*) [17], based on the Rose program [226]. This program adequately simulates the evolution of sequences under no selective pressure and performs random substitutions, deletions, and insertions along each branch, in proportion to their length [16]. For each

alignment size 1000, 5000 and 10000, fifty different multiple sequence alignment were generated (A_1 to A_{50}). Then, we removed all the gaps to obtain 50 sets of unaligned sequences (U_1 to U_{50}) for each alignment size.

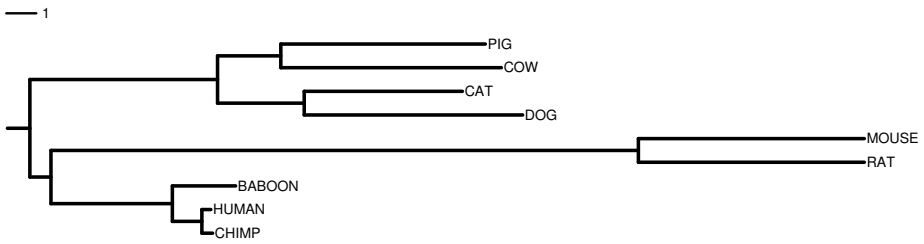


Figure 5–1: The mammalian tree used for the evolution simulation.

Each of the unaligned set of sequences was submitted to following five widely used alignment methods: Clustalw [235], Dialign 2-2 [170], Mafft [129], Mavid [25] and Muscle [71]. We computed a base-per-base alignment accuracy by comparing the obtained alignments to the original ones (see Figure 5–2). It appears that two types of alignment accuracies were generated (good quality with score > 0.8 and poor quality with score < 0.6).

Once the multiple sequence alignments obtained from the alignment programs, they were used to infer phylogenetic trees. Here, we used the distance method bionj [93] (an extension of the popular neighbor joining [205]); the parsimony method Dnapars from PHYLIP [85]; the maximum likelihood methods Dnaml (from PHYLIP) and PHYML [101]; the quartet puzzling method Puzzle [227]. The HKY evolutionary model [105] was used for methods requiring evolutionary model. This is also the model used for the sequence evolution simulation. Each phylogenetic tree obtained was then compared to the true phylogeny using the Robinson and Foulds topological distance [201]. The Robinson and Foulds distance between two phylogenetic trees is the minimum number of operations, consisting of merging and splitting internal nodes, that are necessary to transform one tree into another [201]. This distance is

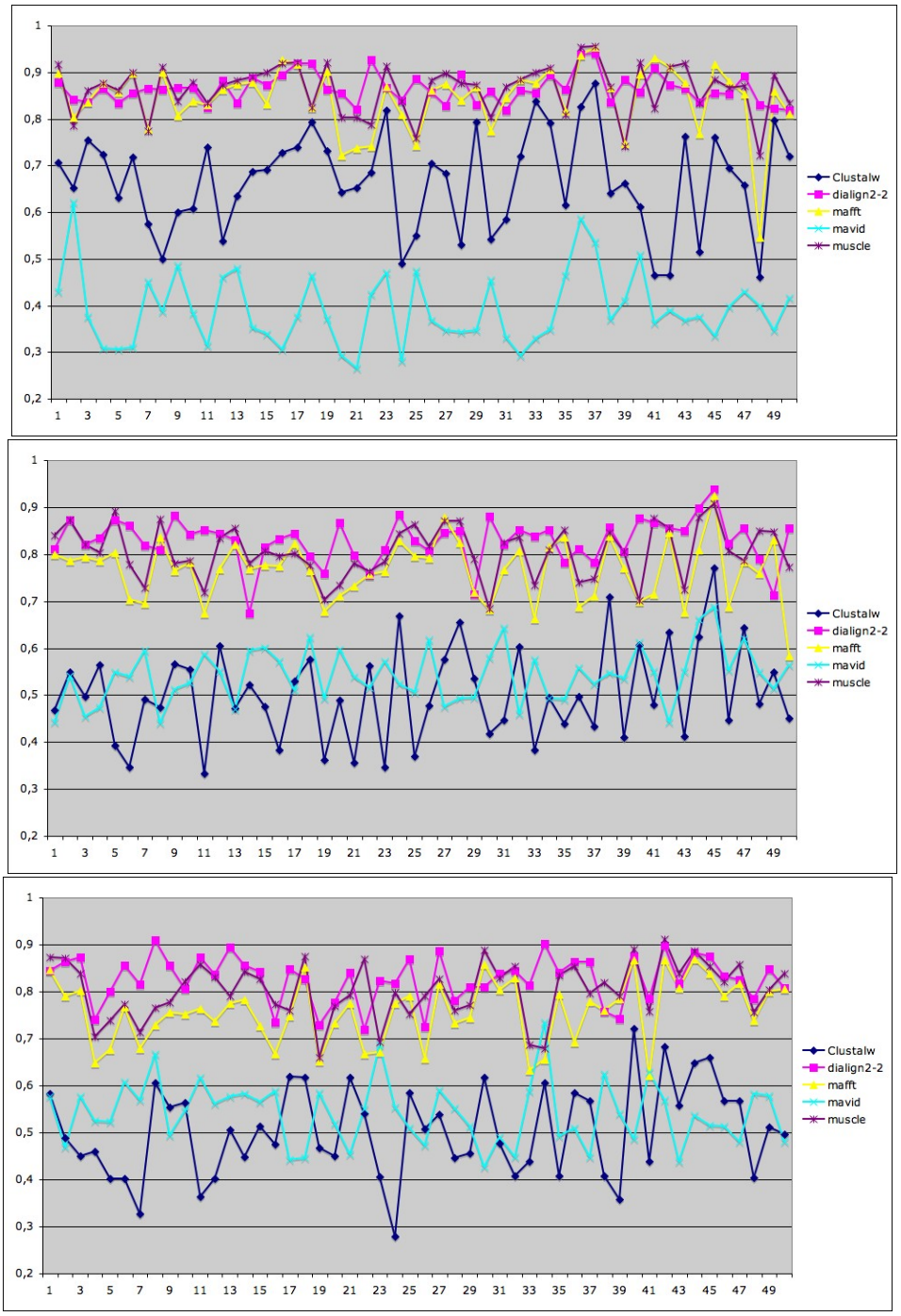


Figure 5–2: The accuracy of the different multiple sequence alignment methods for the alignment size 1000, 5000 and 10000 (from top). The x-axis corresponds to the iteration number. The y-axis corresponds to the base-per-base alignment accuracy.

reported as percentage of its maximum value ($2n$-6 for a phylogeny with n leaves). The lower this value is, the closer the obtained tree to the true tree. Once all trees are obtained, the indel likelihood score is computed for each combination of alignment and phylogenetic tree.

5.6 Results and conclusion

To verify how indel likelihood score correlate with both accuracies of multiple sequence alignments and phylogenetic trees, we took, for each simulation dataset and each alignment size, the alignment-tree pair with the best indel likelihood score. Then, we verified if the associated multiple sequence alignment and phylogenetic trees correspond to those giving respectively the best base-per-base alignment score and the best Robinson and Foulds topological distance. The results presented in Figure 5–3 show that in more than 80% of the cases the best indel score corresponds to the best multiple sequence alignment and the best phylogenetic tree. Despite the fact that in several cases the indel scores does not come from the most accurate multiple sequence alignment and phylogenetic tree, the given results are near the optimal one (less than two Robinson and Foulds operations and between the five three best alignment score).

For comparing alignments of similar size and phylogenetic trees with similar number of taxa, the indel likelihood score can be a good estimate of the agreement between the phylogenetic tree and the multiple sequence alignment. But, it cannot be used as an overall criterion of comparison between non similar data, due to the fact the defined indel likelihood score is dependant of the size of the data. However such a score, should be used in procedure of joint inference of phylogenetic tree and multiple sequence alignment or refining one of them. In this case different dataset size issue is not present. Even though the results in Figure 5–3 are interesting, one might consider the nucleotide

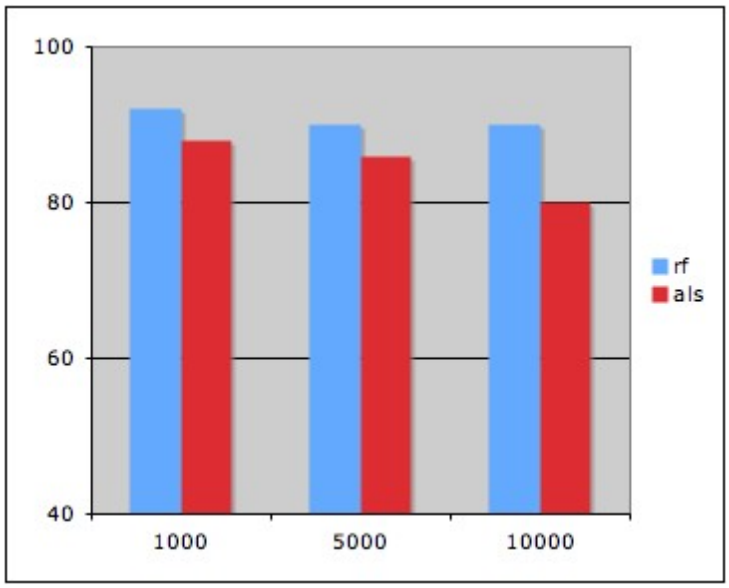


Figure 5–3: Fraction of the simulated datasets where the alignment tree with the best indel likelihood score corresponds to the best alignment (measured by the base-per-base similarity to the correct alignment), identified by *als*, and the best tree (measured by the Robinson and Foulds topological distance to the true tree), identified by *rf*. The results are presented for alignment size 1000, 5000 and 10000.

substitution issue. This issue can be solve by computing the substitution likelihood for all predicted ancestral indel assignment. Even though the latter will require a huge computation complexity, we plan in the future to find an efficient way of combining both. In a different context, the indel likelihood score can be used as a measure of phylogenetic tree method accuracy in place of the widely used Robinson and Foulds topological distance [201] or as alignment method accuracy in various simulation studies. In the case of the Robinson Foulds distance, the original tree of comparison is required for the distance computatio, however it is not needed for the indel score compution. The fact that indel likelihood score uses the branch lengths and the direct association with the alignment will constitute a great advantage. In the future, we plan using the defined score to improve phylogenetic trees through local tree rearrangement such as Nearest Neighbor Interchange and numerical optimization

of branch length. We also plan on refining multiple sequence alignments by making a set of local changes such that the score improves. Finally, all the simulated data used in this study is available at the following URL: <http:// ancestors.bioinfo.uqam.ca/phdDiallo/IndelScore/>.

CHAPTER 6
Étude de Classification des Bactériophages

6.1 Preface

This chapter presents an application of the developed ancestral reconstruction framework to reconstruct the evolutionary history of phages. One major challenge is the presence of eventual horizontal gene transfer and partial data. Here we reconstruct the ancestral protein sequences for Viral orthologous groups and place them in the consensus phylogenetic tree. The text presented in this chapter is taken from Diallo et al. [63], in preparation.

6.2 Résumé

Les bactériophages constituent l'un des groupes d'organismes les plus abondants dans la biosphère. Leur recensement est toujours en cours et les taxonomies proposées sont nombreuses et diverses. Cependant la difficulté intrinsèque, due à la diversité du mode d'évolution et à la complexité de l'écosystème des sujets, est telle qu'une classification exhaustive et convergente reste à établir. Dans cet article, nous présentons une nouvelle approche de l'étude de la classification des bactériophages. Cette approche originale combine à la fois les méthodes de détection des transferts horizontaux de gènes et de reconstruction de séquences ancestrales.

Mots clés: Classification arborescente, Inférence phylogénétique, Transferts horizontaux de gènes, Reconstruction de séquences ancestrales.

6.3 Abstract.

Phages are one of the most present groups of organisms in the biosphere. Their identification continues and their taxonomies are divergent. However,

due to their evolution mode and the complexity of their species ecosystem, their classification is not complete. Here, we present a new approach to the phages classification that combines the methods of horizontal gene transfer detection and ancestral sequence reconstruction.

KEYWORDS: ANCESTRAL SEQUENCE RECONSTRUCTION, CLASSIFICATION, HORIZONTAL GENE TRANSFER, PHYLOGENETIC INFERENCE.

6.4 Introduction

Les bactériophages (ou phages) sont des virus qui infectent les bactéries et les archéobactéries (ou Archaea). Leur évolution est complexe à cause des mécanismes d'évolution réticulée comprenant le transfert horizontal de gènes (THG) et la recombinaison génétique. Une représentation phylogénétique sous forme de réseau est nécessaire pour interpréter l'histoire d'évolution des bactériophages [148]. Par ailleurs, la classification de ces micro-organismes présente intrinsèquement d'autres difficultés dues, d'une part, à la non-conservation de gènes au cours de leur évolution [203], et d'autre part, à la diversité des tailles de leurs génomes [148]. Il existe plusieurs classifications des bactériophages (e.g. [123, 29]). L'approche de classification, adoptée au cours des dernières décennies, est basée sur les critères de morphologie et d'homologie des ADN développés pour les phages (e.g. L. lactis, [123]). La grande majorité des phages a été classée en trois principaux groupes : *936*, *c2* et *P335*. Ainsi, la plupart des études des phages tiennent compte de l'existence de ces groupes. Cependant, plusieurs travaux récents sur l'analyse comparative d'un nombre croissant de séquences génomiques et de l'émergence récurrente de nouveaux phages virulents imposent de facto une révision du mode de classification [224, 199, 50]. Dans cet article, nous proposons une approche en trois phases visant à établir la classification des bactériophages : l'inférence phylogénétique, la détection de transferts horizontaux de gènes [160] et la

reconstruction de séquences protéiques ancestrales [62, 58, 15]. L'arbre phylogénétique des bactériophages a été reconstruit à partir d'une matrice binaire correspondant au contenu en gène de ces organismes. Les THG ont été identifiés pour déterminer le réseau phylogénétique correspondant [160] [Makarenkov et al, 2006]. Enfin, les séquences protéiques ancestrales ont été reconstruites. Cette reconstruction ancestrale permet de mener différentes études approfondies, notamment celle concernant les origines des fonctions protéiques de ces micro-organismes.

6.5 Données Sur Les Bactériophages

6.5.1 Classifications existantes

Les bactériophages sont des micro-organismes très présents dans l'univers. Depuis le séquenage récent de nombreux micro-organismes, les estimations indiquent une population globale de bactériophages de l'ordre de 10^{30}, ce qui représente la forme de vie la plus abondante sur Terre [111]. De nombreux échanges de gènes et des réarrangements de séquences à travers les recombinaisons homologues soumettent les bactériophages à une évolution réticulée [144]. Le THG consiste en un échange direct de matériel génétique d'une lignée à une autre [66]. Bactéries et Archaea ont développé des mécanismes sophistiqués pour acquérir rapidement de nouveaux gènes à l'aide du THG. Le Comité International sur la Taxonomie des Virus (International Committee on the Taxonomy of Viruses ou ICTV) [29] propose une version de taxonomie de ces micro-organismes. Mais la difficulté spécifique due à la diversité du mode d'évolution réticulée et à la complexité de l'écosystème des sujets, est telle qu'une classification exhaustive et convergente n'est pas encore disponible. Par exemple, suivant les classifications établies, la majorité des phages attaquant les bactéries du lait (Lactococcal lactis) appartiendraient à l'un des trois principaux regroupements, notamment, 936, c2 et

P335 [50]. Or, plusieurs travaux récents sur l'analyse comparative des phages semblent démontrer des incohérences dans ces regroupements [224, 223, 199, 50]. Les protéines responsables d'une même fonction dans différents organismes peuvent soit provenir d'un ancêtre commun ou d'une acquisition de gènes indépendante et spécifique à chaque lignée d'espèces [223]. Ainsi, les classifications des phages devraient également étudier le processus d'apparition des fonctions protéiques et l'évolution réticulée de certains gènes.

6.5.2 Données VOG

La banque de données GenBank, hébergée sur le site du National Center for Biotehcnology Information (NCBI) dispose d'une base de données de groupements relatifs aux protéines virales. Cette ressource, nommée Viral COG Clusters of Orthologous Groups (VOG) [5], fournit des molécules standard pour la recherche génomique virale. Les données disponibles proviennent de génomes complets présents dans GenBank. Les données VOG sont des séquences de protéines regroupées de manière prédéfinie en famille selon la fonction protéique à laquelle elles sont associées. Un VOG peut comprendre des séquences de plusieurs espèces différentes. Le contenu informationnel des VOG est utilisé pour améliorer l'annotation fonctionnelle des nouvelles protéines.

L'étude phylogénétique des bactériophages présente une double difficulté en raison de la grande variabilité à la fois de la composition génétique et de la taille des génomes. La première difficulté découle de la grande divergence des séquences protéiques [203]. La seconde difficulté est due aux tailles très disparates de génomes, qui sont d'ordre 2 de magnitude (le nombre de gènes codant en protéines varie de 8 à 381), en comparaison aux procaryotes (de $\sim$400 à $\sim$7 000 gènes) et aux eucaryotes (de $\sim$4 000 à $\sim$60 000 gènes), qui sont d'ordre 1 de magnitude [148]. Bien que la meilleure faon de normaliser les

génomes de ces micro-organismes en vue d'inférer leur histoire d'évolution reste un débat ouvert [168], la tendance actuelle est de combiner l'étude d'évolution du contenu en gènes et l'analyse des alignements de chacune des protéines qui se retrouvent dans les génomes de plusieurs phages [148]. Grâce aux VOG, les regroupements des protéines orthologues apportent des données nécessaires pour résoudre la première partie de notre problème. La méthode de normalisation de l'hétérogénéité de tailles de génomes utilisée sera présentée dans la section suivante. Dans le cadre de cette étude, 163 génomes complets de bactériophages issus de 9 familles différentes, dont une avec des annotations partielles (unclassified), ont été obtenus à partir de GenBank. Les séquences de ces génomes sont distribuées dans 602 regroupements VOG.

6.6 Reconstruction de la phylogénie des bactériophages

La plate-forme d'inférence phylogénétique utilisée prend en entrée les différents groupes de VOG et les séquences protéiques associées à chaque groupe. Elle produit, en premier, un arbre phylogénétique d'espèces qui présente la première hypothèse classique sur l'évolution des bactériophages. Cette présentation en arbre ne tient pas comptes des transferts horizontaux de gènes. La plate-forme produit en second, les arbres de gènes (i.e. des protéines) individuels qui représentent l'évolution de chacun des gènes considérés dans les VOG.

6.6.1 Construction de l'arbre phylogénétique d'espèces

Pour construire la classification arborescente des bactériophages, une matrice binaire de présence et d'absence de gènes (i.e. de regroupements VOG) dans chacun des 163 phages a été composée (Figure 1). Ainsi, la matrice obtenue contient 163 lignes et 602 colonnes qui représentent respectivement les phages et les VOG. Les méthodes d'inférence phylogénétique utilisant une

approche de distance et une approche bayésienne ont été appliquées pour reconstruire l'arbre (voir Bartélemy et Guénoche [1988] pour plus de détails sur les techniques d'inférence d'arbres phylogénétiques). Des tests de robustesse ont été effectués pour mesurer le taux des regroupements d'espèces présents dans l'arbre obtenu.

Méthode de distance. Comme requis par les méthodes de distance, une matrice de dissimilarités inter-génomiques a été initialement calculée. Plusieurs types de distances ont été récemment utilisés pour mesurer la distance entre les génomes : le coefficient de corrélation standard [97], le coefficient de Jaccard [97], le coefficient de Maryland Bridge [168] et la Moyenne Pondérée [68]. Dans cette étude, nous avons testé ces différents coefficients. Les résultats obtenus étaient très similaires, compte tenu qu'il n'y a pas d'ordre a priori dans les regroupements VOG. Dans notre étude, le coefficient de Jaccard a été utilisé. La méthode de Neighbor Joining (NJ) [205] a permis d'inférer l'arbre phylogénétique d'espèces. Un test de bootstrap a été réalisé pour évaluer la stabilité des groupes présents dans les topologies en fonction de différents échantillons de la matrice binaire. Les différents échantillons aléatoires ont été obtenus à l'aide du programme SeqBoot inclus dans le paquet PHYLIP [85]. Dans cette étude 100 échantillons ont été retenus. À la suite de l'inférence phylogénétique de chacun des échantillons, un arbre de consensus a été inféré pour chacune des approches. Le programme Consense inclus également dans le paquet PHYLIP a servi à générer l'arbre de consensus par la règle de majorité étendue (50%).

Méthode bayésienne. L'inférence bayésienne produit un arbre phylogénétique à partir de la distribution a posteriori des topologies d'arbres. Elle évalue l'espace de solutions au moyen des chanes de Markov. Dans cette étude, le logiciel MrBayes [119, 118] a été utilisé avec 2 millions de générations

échantillonnées à toutes les 100 générations, 4 chanes et 2 exécutions indépendantes, créant ainsi 20 000 arbres. Un arbre de consensus a été inféré à partir des 1000 derniers arbres les plus stables (i.e. générations stationnaires). Le programme Consense a servi à construire l'arbre de consensus par la règle de majorité étendue (50%).

6.6.2 Inférence des arbres de gènes

Un arbre de gène a été inféré pour chaque groupe VOG (Figure 1). Les séquences protéiques associées à un VOG donné ont été alignés en utilisant ClustalW [235]. Le programme MrBayes a été utilisé pour inférer les arbres de gène pour les 602 alignements de séquences de VOG.

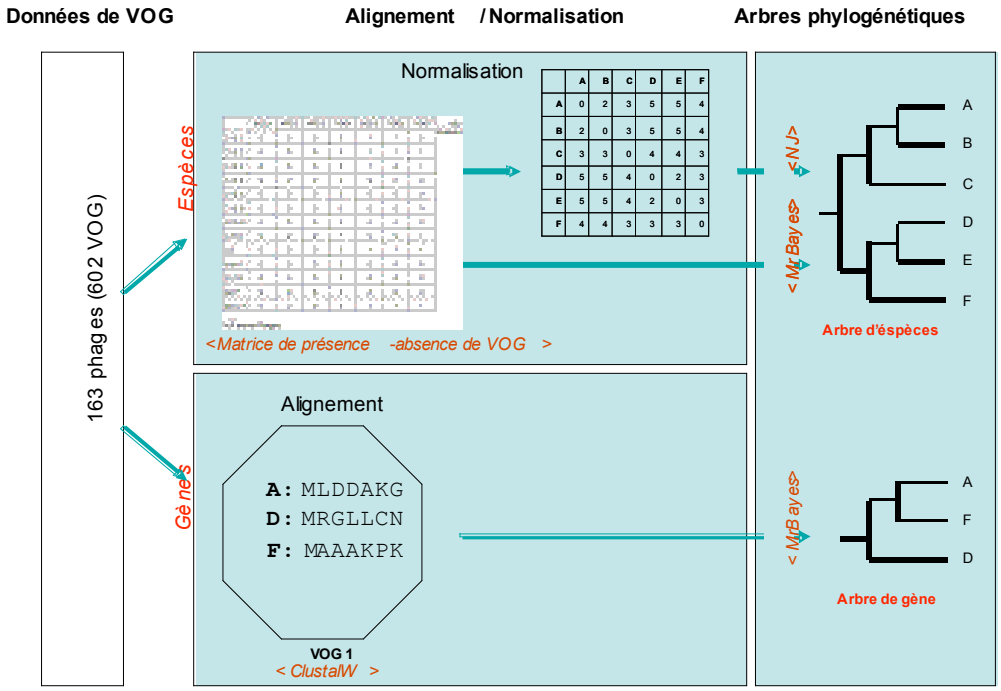


Figure 6–1: Construction des arbres phylogénétiques d'espèces et de gènes; pour les arbres d'espèces, une matrice binaire a été obtenue en fonction de la présence ou de l'absence de phage dans chaque VOG. Les arbres de gènes ont été inférés à partir des séquences protéiques alignées.

6.7 Détection des THG

La détection des THG a été effectuée, en utilisant le programme HGT Detection du package T-Rex [155], suivant une version améliorée de l'algorithme de réconciliation topologique entre l'arbre de gènes et l'arbre d'espèces [160]. HGT Detection (Cf. le site <www.trex.uqam.ca>) prend en entrée un arbre d'espèces et un arbre de gènes pour le même ensemble d'espèces. Les THG sont ainsi calculés, en indiquant en sortie l'origine et la destination pour chacun des transferts inférés. Les principales étapes de l'algorithme heuristique pour identifier des THG sont les suivantes :

Pas préliminaire

Inférer les arbres phylogénétiques d'espèces (i.e. arbre de contenu en gène dans notre cas) et celui de gène (l'arbre du VOG considéré dans notre cas), notés respectivement T et T'. L'arbre T est un arbre réduit de l'arbre complet construit pour 163 bactériophages et contenant seulement les phages présents dans le VOG considéré. Les deux arbres doivent être enracinés. S'il existe dans T et T' des sous-arbres identiques ayant au moins 2 feuilles, réduire la taille du problème en remplaant dans T et T' les sous-arbres identiques par les mêmes éléments auxiliaires.

Pas $1 \ldots k$

Tester tous les THG possibles entre les paires d'arêtes dans l'arbre T_{k-1} ($T_{k-1} = T$ au Pas 1) à l'exception des transferts entre les arêtes adjacentes et ceux qui violent les contraintes d'évolution (pour plus de détails voir [189]). Choisir en tant que THG optimal, le déplacement d'un sous-arbre dans T_{k-1} qui minimise la valeur de la distance topologique de Robinson et Foulds [201] entre l'arbre obtenu après le déplacement de ce sous-arbre et de son greffage sur

une nouvelle arête, i.e. l'arbre T_k, et l'arbre de gène T'. Réduire ensuite la taille du problème en remplaant des sous-arbres identiques, ayant au moins 2 feuilles, dans l'arbre transformé T_k et l'arbre de gène T'. Dans la liste des THG retrouvés rechercher et éliminer les THG inutiles en utilisant une procédure de programmation dynamique de parcours en arrière. Un transfert inutile est celui dont l'élimination ne change pas la topologie de l'arbre T_k.

Conditions d'arrêt et complexité algorithmique

L'algorithme s'arrête quand la distance de Robinson et Foulds devient égale à 0 ou quand aucun autre déplacement de sous-arbres n'est possible suite à des contraintes biologiques. Théoriquement, une telle procédure requiert $O(kn^4)$ d'opérations pour prédire k transferts dans un arbre phylogénétique à n feuilles. Cependant, due à des réductions inévitables des arbres d'espèces et de gènes, la complexité pratique de cet algorithme est plutôt $O(kn^3)$.

6.8 Reconstruction des séquences protéiques ancestrales

La reconstruction ancestrale permet d'étudier l'évolution des espèces, la sélection adaptative et la divergence fonctionnelle [135]. Elle tient compte de la recréation de protéines et de l'évolution d'ADN en laboratoire de sorte qu'elles puissent être étudiées directement [33]. De plus, la reconstruction ancestrale de protéines peut mener aux découvertes de nouvelles fonctions biochimiques qui ont été perdues au cours de l'évolution [124]. Les séquences de protéines renferment des informations sur leurs passés historiques [191]. Dans cette étude, nous nous sommes intéressés en particulier aux fonctions protéiques et à la recherche de séquences ancestrales des VOG. Ces séquences ancestrales faciliteraient l'analyse de similitude structurale des acides aminés en présentant des séquences représentatives de groupes de phages. Cette étude permet de réduire

la complexité des analyses. Les séquences protéiques ancestrales et leur probabilité a posteriori au niveau de chaque caractère sont prédites. Ces protéines ancestrales pourraient servir comme représentants de familles de bactériophage lors de différentes analyses de génomique comparée. La reconstruction des séquences protéiques ancestrales s'effectue en deux étapes: reconstruction des ancêtres et représentation des séquences obtenues dans l'arbre d'espèces déjà construit.

6.8.1 Reconstruction des séquences ancestrales

Au préalable, les séquences protéiques de chaque VOG ont été alignées en utilisant le programme d'alignement de séquences multiples ClustalW [235]. Les arbres phylogénétiques représentant l'histoire d'évolution de chacun des VOG ont été reconstruits à l'aide du programme MrBayes [119, 118] . L'arbre de consensus a été inféré, puis enraciné, en utilisant la technique du point médian (midpoint). Étant donné un alignement de séquences de régions orthologues et un arbre phylogénétique, la reconstruction de séquences ancestrales consiste à l'inférence pour chaque nœud interne de l'arbre phylogénétique, de la séquence génomique correspondante. Cette inférence s'effectue en deux étapes: la reconstruction du scénario d'insertion et de délétion (i.e. indel) le plus vraisemblable, et l'inférence des acides aminés à chaque position des ancêtres où la présence d'un caractère a été prédite. Ces deux étapes sont réalisées respectivement par les algorithmes de Diallo et al. [62, 58] et Felsenstein [82] qui sont implantés dans le programme Ancestor disponible à l'URL suivant : <www.mcb.mcgill.ca/∼banire/ancestor>.

6.8.2 Représentation des séquences ancestrales

Les séquences ancestrales obtenues sont représentées sur l'arbre d'espèce (Figure 3a) en utilisant la technique de l'ancêtre commun le plus proche (ACP). Ainsi chaque VOG est associé à une séquence ancestrale située au

nœud ancestral minimal commun. Il est important de noter que dans la présente étude, l'ordre des VOG n'a pas été pris en compte. Il serait intéressant dans un travail futur de trier les gènes pour déterminer l'ordre exact dans les séquences ancestrales [21]. Cette représentation permet d'identifier au cours de l'évolution les diverses apparitions de nouvelles fonctions. Ainsi pour chaque VOG, la séquence ancestrale a été inférée. Un exemple de ces résultats présentés dans la section suivante concerne la famille des phages attaquant les bactéries du lait, L. lactis (sous-section 6.3).

6.9 Résultats

6.9.1 Reconstruction de la phylogénie des bactériophages

Les arbres phylogénétiques d'espèces inférés par NJ et MrBayes étaient très similaires au niveau des regroupements d'espèces (avec de meilleurs scores de robustesse au niveau des groupes retrouvés par MrBayes). Ceci converge avec les études antérieures de comparaison d'inférence phylogénétique préconisant une meilleure précision pour des méthodes bayésiennes comparativement aux méthodes de distance [140]. Ainsi, la Figure 2 montre l'arbre phylogénétique de bactériophages avec les différentes statistiques obtenues pour la méthode bayésienne. Globalement, l'arbre phylogénétique d'espèces incorpore un grand nombre de signaux phylogénétiques : au total, 116 phages, c-à-d 71% des génomes étudiés, ont été classés dans 22 groupes avec des scores de probabilités a posteriori supérieur à 50%. Ces groupes robustes contiennent entre 3 et 10 phages, avec une taille moyenne de clades de 6 espèces. Plusieurs familles d'espèces, 12 sur 22 groupes, référencées par ICTV ont été retrouvées par notre analyse : Siphoviridae (groupes 1, 2, 6, 8, 9, 10, 13, 22), Podoviridae (groupes 14, 20, 21) et Myoviridae (groupe 4). Cependant plusieurs clades demeurent non résolus. Cela est dû à l'absence d'information convergente au niveau du contenu en gène, traduit par la présence de différentes topologies

associées à ces organismes parmi les arbres générés par MrBayes. Par ailleurs, on constate que la plupart de ces clades font partie des espèces partiellement annotées (unclassified).

6.9.2 Détection des THG

Au niveau des transferts, nous avons calculé les statistiques globales des THG intra (Within) et inter (In/Out) groupes. Plusieurs points sont remarquables : (a) les groupes 2, 5, 6, 12 à 16, 21 et 22 ont un nombre de transferts intra-groupes supérieur à ceux d'inter-groupes, alors que le reste des groupes a une tendance inverse, à l'exception cependant des groupes 4 et 9 qui n'ont pas de transferts intra-groupes ; (b) les groupes 1, 5, 6, 7, 10, 12, 14, 16, 21 et 22 en donnent plus qu'ils en reoivent, et inversement pour le reste, à l'exception du groupe 11 qui ne donne ni reoit de transferts, et du groupe 8 qui en donne autant qu'il en reoit ; (c) les groupes qui en donnent beaucoup plus que la moyenne (informations non représentées sur la Figure 2) sont les suivants : le groupe 1 au groupe 8 (19 transferts), le groupe 17 au groupe 20 (12 transferts), le groupe 20 au groupe 17 (15 transferts) et le groupe 8 au groupe 1 (14 transferts). Les transferts entre les espèces hors groupes (i.e. les clades non résolus discutés plus haut) n'ont pas été comptabilisés dans cette étude.

6.9.3 Reconstruction des séquences protéiques ancestrales

Les résultats de la procédure de reconstruction des séquences protéiques ancestrales sont présentés sous forme d'arbres et de tableaux (Figure 3ab, vue partielle). Ainsi, nous déterminons pour chaque VOG, sa protéine ancestrale et le nœud ancestral correspondant dans l'arbre d'espèces. Ce travail permet d'identifier, à des fins de comparaison de génomes, l'ensemble des fonctions assignées à chaque nœud ancestral de l'arbre d'espèces (Figure 3b, voir les

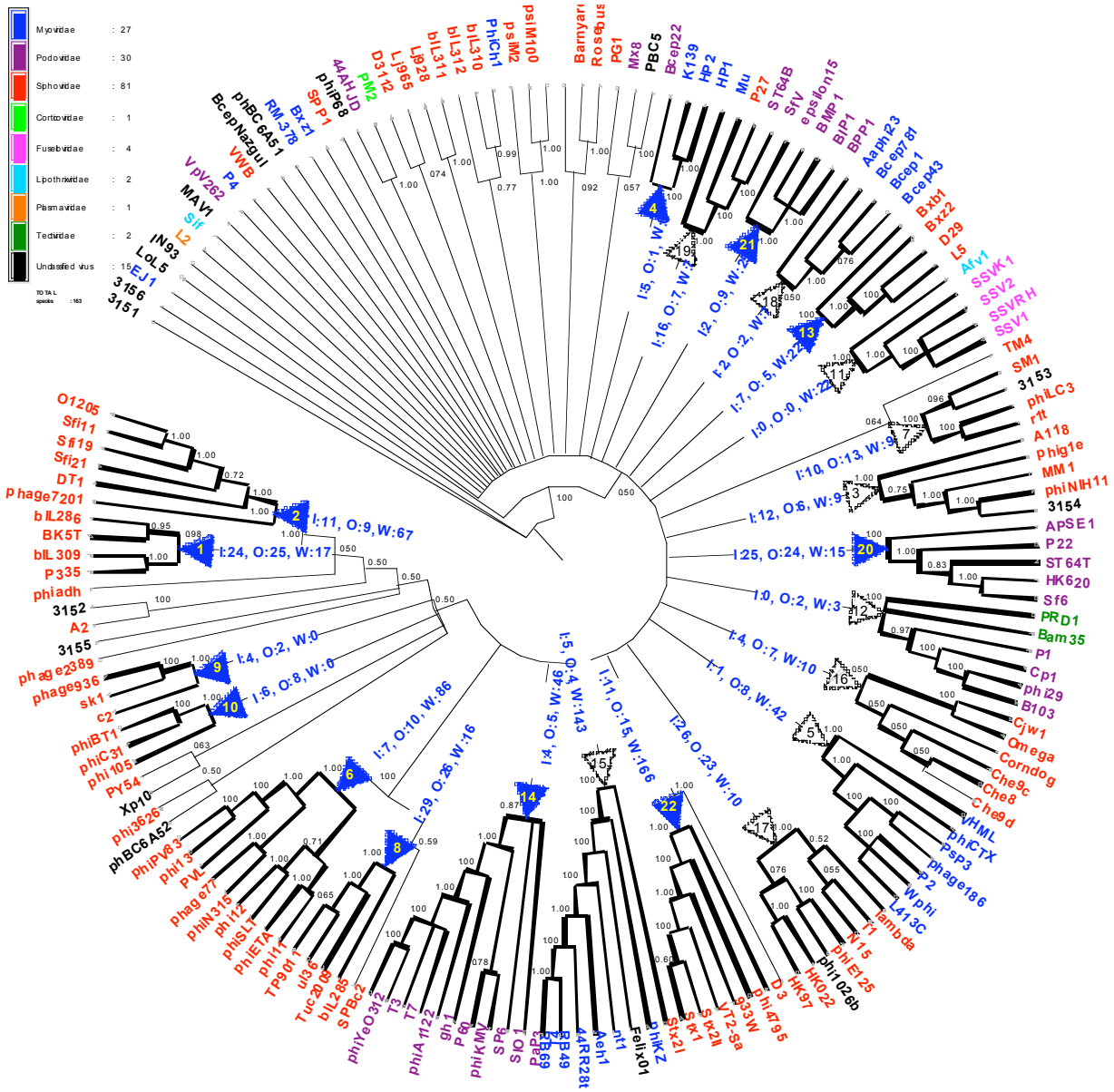


Figure 6–2: Arbre phylogénétique d'espèces inféré par MrBayes [119, 118] . Les scores de robustesse sont indiqués pour les arêtes internes; 12 des 22 groupes identifiés (représentés par des triangles pleins) correspondent aux taxonomies de NCBI/ICTV. Pour chaque groupe, I (In) signifie le nombre de THG entrant dans le groupe, O (Out) le nombre de THG sortant du groupe et W (Within) le nombre de THG à l'intérieur de ce groupe. La figure a été dessinée à l'aide de l'outil de représentation d'arbres iTol (disponible sur : http://itol.embl.de).

Annexes pour les résultats complets). Une prochaine étape serait la comparaison des structures des protéines ancestrales reconstruites à celles des sous-groupes correspondants. Ce travail permettrait de déterminer quels sont les

domaines conservés dans les ancêtres, les fonctions existantes, les variations des séquences protéiques, les fonctions perdues par certains organismes, les fonctions acquises par les organismes de faon indépendante, etc. Ces résultats permettraient également de définir des séquences consensus pour les sous-arbres de phages, ce qui permettrait de réduire la complexité de la comparaison intergroupe de phages lors des analyses comparatives de séquences. À des fins de validation et de comparaison, il est important de mentionner ici qu'un score de prédiction de chaque séquence ancestrale (et des caractères prédits) a été également calculé en utilisant la probabilité à posteriori d'inférence de chaque caractère.

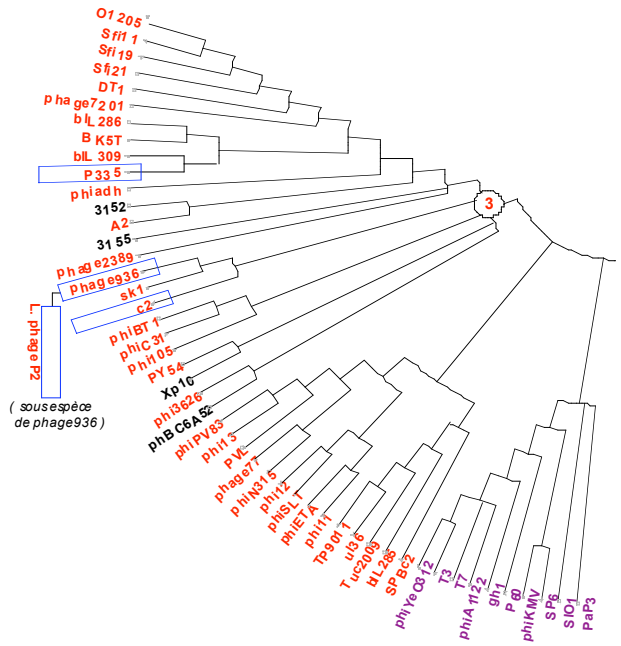
Considérons par exemple le nœud 3 (Figue 3a) qui est le nœud ancestral commun le plus proche des phages 936 (et L. phage P2), c2 et p335. Selon la taxonomie de référence d'ICTV, les phages attaquant les bactéries de L. lactis sont membres de l'ordre des Caudovirales qui regroupe trois familles : Myoviridae, Siphoviridae et Podoviridae. Tous les phages attaquant les bactéries de L. lactis connus sont principalement membres de la famille Siphoviridae [50] (commenant, en suivant l'ordre circulaire, par l'espèce O1205 et se termine par l'espèce SPBc2; Figure 3a) et quelques espèces de la famille Podoviridae (commenant par l'espèce phyYeO312 et se termine par l'espèce *PaP3*). Figure 3a présente un exemple d'un sous-arbre de l'arbre d'espèces complet (Figure 2) comprenant les phages attaquant les bactéries de lait. Le nœud 3 est l'ancêtre commun le plus proche des phages *P335*, *Phage936*, *c2* (et *L. phage P2*). Deux séquences protéiques ancestrales inférées pour ces organismes, leur fonction et les VOGs correspondants sont reportés dans la table (Figure 3b).

6.10 Conclusion

L'approche présentée ici combine à la fois les méthodes de détection de transferts horizontaux de gènes et de reconstruction de séquences ancestrales

pour proposer une autre hypothèse sur la classification des bactériophages. Les résultats obtenus apportent des informations additionnelles qui visent à mieux comprendre l'histoire d'évolution de ces micro-organismes. En effet, l'issue de cette étude a permis de : (a) fournir une classification des bactériophages tenant compte de l'évolution réticulée, (b) fournir des statistiques sur les différents transferts horizontaux inter et intra-groupes, (c) générer des séquences ancestrales des phages et identifier leur origine dans l'histoire évolutive de ces organismes. Le dernier point permet aussi d'identifier des patrons communs aux groupes de séquences. Cependant, dans cette étude, nous avons occulté plusieurs problèmes dont ceux liés à l'exactitude des alignements obtenus, les scénarios de reconstructions ancestrales alternatifs ainsi que le problème lié à l'ordre des VOG dans les différents génomes. Les statistiques complètes concernant la classification des bactériophages sont disponibles à l'URL suivant : <http://www.info2.uqam.ca/∼makarenv/Annexe_SFC2007.pdf>.

a)



b)

Noeud interne	Fonction	VOG	AncSeq
3	[T] Putative DNA binding protein	vogp0362	MAN
3	[U] Uncharacterized	vogp0126	

Figure 6–3: (a) Sous-arbre de l'arbre d'espèces complet (Figure 2) comprenant les membres des familles Siphoviridae et Podoviridae. Le nœud 3 est l'ancêtre commun le plus proche des phages attaquant les bactéries de L. lactis, i.e. les organismes P335, Phage936, c2 (et L. phage P2). (b) Deux séquences protéiques ancestrales inférées pour ces organismes, leur fonction et les VOGs correspondants.

CHAPTER 7
A whole genome study and identification of specific carcinogenic regions of the Human Papilloma Viruses

7.1 Preface

This chapter contains a large evolutionary study including indel analyses of all the complete sequenced Human Papilloma Viruses. Here we analyze how indel distribution obtained from our developed framework can be related to the carcinogenic structure of the phylogenetic tree. It also presents an algorithm to identify specific carcinogenic regions with respect to their evolutionary properties (small scale mutations). The text presented in this chapter is taken from Diallo et al. 2009 [55], accepted for publication into the journal of Computational Biology.

7.2 abstract

In this article, we undertake a study of the evolution of Human Papillomaviruses (HPV), whose potential to cause cervical cancer is well known. First, we found that the existing HPV groups are monophyletic and that the high-risk of carcinogenicity taxa are usually clustered together. Then, we present a new algorithm for analyzing the information content of multiple sequence alignments in relation to epidemiologic carcinogenicity data to identify regions that would warrant additional experimental analyses. The new algorithm is based on a sliding window procedure and a p-value computation to identify genomic regions that are specific to HPVs causing disease. Examination of the genomes of 83 HPVs allowed us to identify specific regions that might be influenced by insertions, deletions, or simply by point mutations, and that may be of interest for further analyses.

7.3 Introduction

Human papillomaviruses (HPV) have a causal role in cervical cancer with almost half a million new cases identified each year [3, 19, 175]. The HPV genomic diversity is well known [4]. About one hundred HPV types are identified, and the whole genomes of more than eighty of them are sequenced (see the latest Universal Virus Database report by International Committee on Taxonomy of Viruses (ICTV)). A typical HPV genome is a double-stranded, circular DNA genome of size close to 8 Kbp, with complex evolutionary relationships and a small set of genes. In general, the E5, E6, and E7 genes modulate the transformation process, the two regulatory proteins, E1 and E2, modulate transcription and replication, and the two structural proteins L1 and L2 compose the viral capsid. Protein E4 has an unclear function in the HPV life cycle, however, several studies indicate that it could facilitate the viral genome replication and the activation of viral late functions [253], and it could also be responsible for virus assembly [193]. A HPV is considered to belong to a new HPV type if both its complete genome has been cloned and the DNA sequence of the gene L1 differs by more than 10% from the closest known HPV type. The comparison of HPV genomes, conducted by ICTV, is based on nucleotide substitutions only [177, 46]. Older HPV classifications were built according to their higher or lower risk of cutaneous or mucosal diseases. Most of the HPV studies were based on single gene (usually E6 or E7) analyses. The latter genes are predominantly linked to cancer due to the binding of their products to the p53 tumor suppressor protein and the retinoblastoma gene product pRb [242]. To define carcinogenic types, we used epidemiologic data from a large international survey on HPVs in cervical cancer and from a multicenter case-control study conducted on 3,607 women with incident, histologically confirmed cervical cancer recruited in 25 countries

[177, 176]. HPV DNA detection and typing in cervical cells or biopsies were centrally done using PCR assays which attest for the quality of the study [177]. More than 89% of patients them had squamous cell carcinoma (i.e. Squam cancer) and about 5% had adenosquamous carcinoma (i.e. Adeno cancer) see Table 7–1 adapted from [177]. More than half of the infection cases are due to the types 16 and 18 of HPV, which are thus referred to as high-risk HPVs [32].

Table 7–1: Distribution of carcinogenic HPVs for the Squam and Adeno types of cancer. Complete genomic sequence data is not available yet for HPVs-35, HR, 68, and X.

	Squamous cell carcinoma		*Adenocarcinoma and adenosquamous carcinoma*	
HPV types	**Number**	**% positive**	**Number**	**% positive**
HPV-16	1,452	54.38	77	41.62
HPV-18	301	11.27	69	37.30
HPV-45	139	5.21	11	5.95
HPV-31	102	3.82	2	1.08
HPV-52	60	2.25		
HPV-33	55	2.06	1	0.54
HPV-58	46	1.72	1	0.54
HPV-56	29	1.09		
HPV-59	28	1.05	4	2.16
HPV-39	22	0.82	1	0.54
HPV-51	20	0.75	1	0.54
HPV-73	13	0.49		
HPV-82	7	0.26		
HPV-26	6	0.22		
HPV-66	5	0.19		
HPV-6	2	0.07		
HPV-11	2	0.07		
HPV-53	1	0.04		
HPV-81	1	0.04		
HPV-55	1	0.04		
HPV-83	1	0.04		
Total	**2,293**	**85.89**	**168**	**90.37**

In this paper, we first studied a whole genome phylogenetic classification of the HPV and the insertion and deletion (indel) distribution among HPV lineages leading to the different types of cancer. First, we inferred a phylogenetic tree of 83 HPVs based on whole HPV genomes. We found that the evolution of the L1 gene, used by ICTV to establish the HPV classification, generally reflects the whole genome evolution. Second, we compared the gene trees built for the 8 most important HPV genes (E1, E2, E4, E5, E6, E7, L1 and L2) using the normalized Robinson and Foulds topological distance [201]. Then, we described a new algorithm for analyzing the information content of multiple sequence alignments in order to identify regions that may be responsible for the carcinogenicity. This algorithm is based on a new formula taking into account the sequence similarity among carcinogenic taxa and the sequence dissimilarity between the carcinogenic and non-carcinogenic taxa, computed for a genomic region bounded by the position of the sliding window. To facilitate the identification of relevant regions, we compute p-values for the different regions according to their score obtained with our new formula. Using the new technique we developed, we examined all available genes in 83 HPV genomes and identified the specific genomic regions that would warrant interest for future biological studies.

7.4 Indel analysis of HPV genomes and reconciliation of HPV gene trees

The 83 completely sequenced HPV genomes (all identified by the ICTV) were downloaded and aligned using ClustalW [235], producing an alignment with 10426 columns. The phylogenetic tree of 83 HPVs (Figure 7–1) was inferred using the PHYML program [101] with the HKY substitution model. Bootstrap scores were computed to assess the robustness of the edges using 100 replicates. Most branches obtain support above 80%, but for a better

readability, they are not represented in Figure 7–1. However, they are given in the supplemental materials [1] . As suggested in [242], the bovine PV of type 1 was used as outgroup to root this phylogeny. To the best of our knowledge, the constructed phylogenetic tree is the first whole genome phylogenetic tree of HPVs.

Our analysis revealed the presence of 12 known monophyletic HPV groups that are denoted by numerated nodes, labeled according to the ICTV annotation, in Figure 7–1. The other monophyletic groups obtained were not depicted by numbers. The whole-genome phylogeny obtained usually corresponds to the HPV classification provided by ICTV on the basis of the L1 gene. Most of the dangerous HPVs (see Table 7–1) can be found in the sister subtrees rooted by the nodes 16 and 18.

As carcinogenicity may be introduced into a HPV by an insertion or deletion (indel) of a group of nucleotides, we first addressed the problem of indel distribution in the evolution of HPV. Thus, the most likely indel scenario was inferred using a heuristic method described in [62, 58]. Such a scenario includes the distribution of the predicted indel and base conservation events for all HPV genes. Table 7–2 reports, for each of the 8 main genes of HPV, the total number of conservations, insertions and deletions of nucleotides that occurred during their evolution. Genes E1, L1 and L2 show more than 90% conservation at the nucleotide level, E2, E4 and E6 between 80 and 90%, and E5 and E7 respectively 73% and 59%.

[1] Supplemental materials are available at:
http://ancestors.bioinfo.uqam.ca/articles/JCB2009/supplemental.zip

Table 7–2: For each of the 8 main HPV genes, this table reports the numbers (and average numbers) of Conservations (including substitutions), Insertion and Deletions of nucleotides that occurred during evolution.

Variable/Gene	*Conservation*	*Insertion*	*Deletion*	*Avg. Cons.*	*Avg. Ins.*	*Avg. Del.*
E1	12111	601	2774	0.918	0.003	0.010
E2	13304	306	3460	0.852	0.001	0.022
E4	6318	195	2117	0.851	0.001	0.038
E5	1688	356	503	0.731	0.021	0.031
E6	7323	613	1529	0.890	0.002	0.011
E7	3457	0	1393	0.594	0.000	0.039
L1	9664	314	2751	0.927	0.001	0.010
L2	21716	494	5138	0.923	0.004	0.026

The highest indel frequencies are in the subtrees rooted at node 61 where there are only low risks of carcinogenicity (Figure 7–1). The groups included in the subtree A have low percentage of indels on in each branch. Thus, one could conclude that indel rates could not be related to gained of carcinogecity. However, due to the fact that it is less likely that carcinogenicity has been gained several times independently (through substitutions for instance) in different taxa, one plausble hypothesis is that the organisms of the subtree A inherited their carcinogenicity from their closest common ancestor.

We also carried out an analysis intended to compare the topologies of the gene phylogenies built for the 8 main HPV genes. Thus, we first aligned, using ClustalW [235], the HPV gene sequences, separately for each gene, and inferred 8 gene phylogenies using the PHYML program [101] with the HKY model. In order to measure their degree of difference, we computed the Robinson and Foulds (RF) topological distances between each pair of gene trees [201]. As the number of tree leaves varied from 70 to 83 (due to the non-availability of some gene sequences for a few HPVs), we reduced the size of some trees prior to this pairwise topological comparison and normalized all distances by the largest possible value of the RF distance, which is $2n - 6$ for two binary trees with n leaves. Figure 7–2 shows the results obtained, with RF distances are depicted as stacked rectangles. The results suggest that the trees representing the evolution of the E4 and E5 genes differ the most, on average, from the other gene phylogenies, whereas the phylogeny of E2 reconciles the most the topological differences of this group of gene trees. Two HPV gene phylogenies differ from each other by about 32%, on average. In the future, it might also be interesting to compare the gene trees we obtained using Maximum Likelihood tests such as Shimodaira-Hasegawa [213] or Kishino-Hasegawa [133] and to

assess the confidence of phylogenetic tree selection using program such as CONSEL [214].

These results confirm the hypothesis made in a number of HPV studies (see for instance [179, 243]), that most HPV genes undergo frequent recombination events. Uncritical phylogenetic analyses performed on recombinant

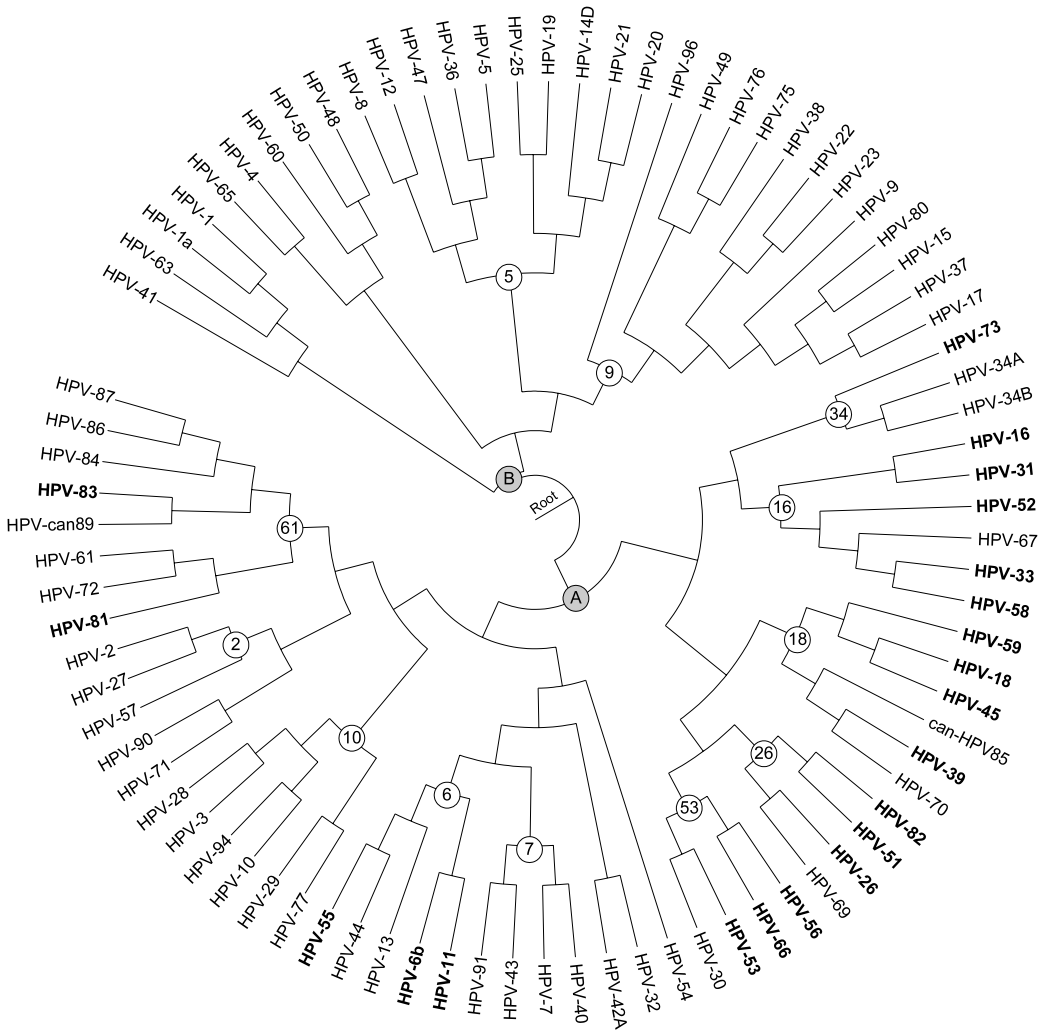


Figure 7–1: Phylogenetic tree of 83 HPVs obtained with PHYML. The 21 carcinogenic HPV are shown in bold. The white nodes identify the existing HPV groups according to the ICTV and NCBI taxonomic classifications; the shaded nodes (A and B) distinguish between the non-carcinogenic and carcinogenic families. Bootstrap scores are above 80% for most of the branches; for a better readability, they are not represented. The HPVs 1 and 34 are present in two copies, (1 and 1a) and (34A and 34B), respectively.

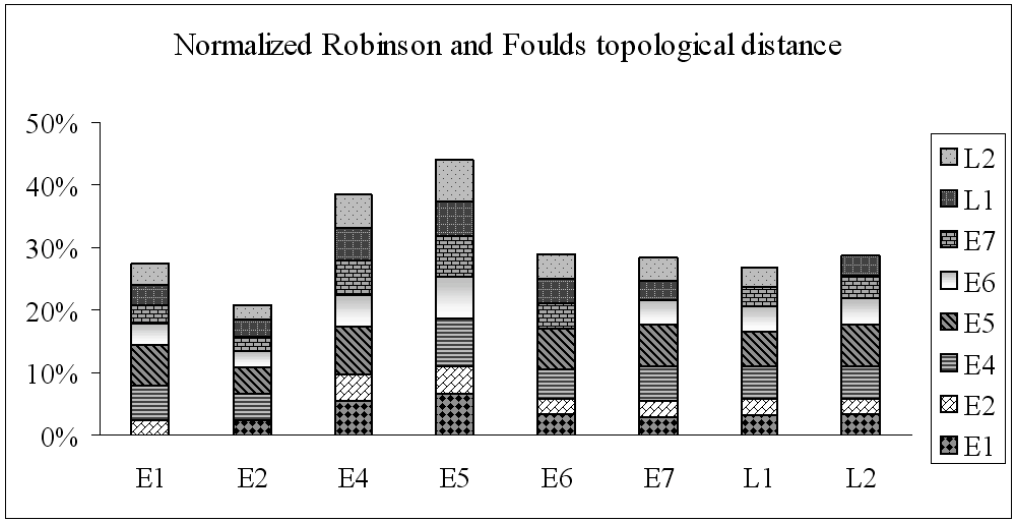


Figure 7–2: Average normalized Robinson and Foulds topological distance for each of the 8 main HPV genes. Each column of the diagram represents a gene and consists of the stacked rectangles whose heights are proportional to the values of the normalized Robinson and Foulds topological distances between the phylogeny of this gene and those represented by the stacked rectangles. The column heights depicts the total average distance. For the sake of presentation the percentage values on the ordinate axis were divided by 7 (which is the number of pairwise comparisons made for each gene tree).

sequences could lead to the impression of novel, relatively isolated branches. Recently, Angulo and Carvajal-Rodriguez (2007) have provided new support to the recent evidence of recombination in HPV. They found that the gene with recombination in most of the groups is L2 but the highest recombination rates were detected in L1 and E6. Gene E7 was recombinant only within the HPV16 type. The authors concluded that this topic deserves further study because recombination is an important evolutionary mechanism that could have a high impact both in pharmacogenomics and for vaccine development.

7.5 Algorithm for the identification of putatively carcinogenic regions

This section describes a new algorithm intended for finding genomic regions that may be responsible for HPV carcinogenicity. The algorithm is based on the hypothesis that sequence regions responsible for cancer are likely to be

more similar among carcinogenic HPVs than between carcinogenic and non-carcinogenic HPVs. The following procedure was adopted. First, 83 available HPV genomes were downloaded and inserted into a relational database along with the clinical information regarding identified HPV types and histological type of cancer occurrences [177, 176]. We constructed three HPV Types Datasets: "High-Risk", containing HPVs16 and 18, "Squamous", containing HPV types responsible for Squamous Cell Carcinoma (HPV-6, 11, 16, 18, 26, 31, 33, 39, 45, 51, 52, 53, 55, 56, 58, 59, 66, 73, 81, 82, 83) and Adeno with types responsible for Adenocarcinoma (HPV-16, 18, 31, 33, 35, 39, 45, 51, 58, 59). See Table 7–1 for more details. HPV types with incomplete genome information or without annotations were excluded from the dataset. As previously, we used the gene sequences aligned separately for each gene.

Then, we scanned all gene sequence alignments using a sliding window of a fixed width (in our experiments the window width ranged from 3 to 20 nucleotides, see Figure 7–3). First, a detailed scan of each gene with increments of 1 nucleotide was performed to identifying the regions with a potential for causing carcinogenicity (the main results are reported in Table 7–3), and called here hit regions. Second, a non-overlapping windows of width 20 nucleotides was carried out for plotting Figures 7–4, 7–7 and 7–8. Three separate analyses were made for the three above-described carcinogenic families: High-Risk, Squamous and Adeno HPVs.

Once the window position is fixed and the taxa are assigned to the sets X (carcinogenic HPVs) and Y (non-carcinogenic HPVs), the hit region identification function, denoted here as Q, can be computed. This function is defined as a difference between the means of the squared distances computed among the sequence fragments (bounded by the sliding window position) of the taxa from the set X and those computed only between the sequence fragments

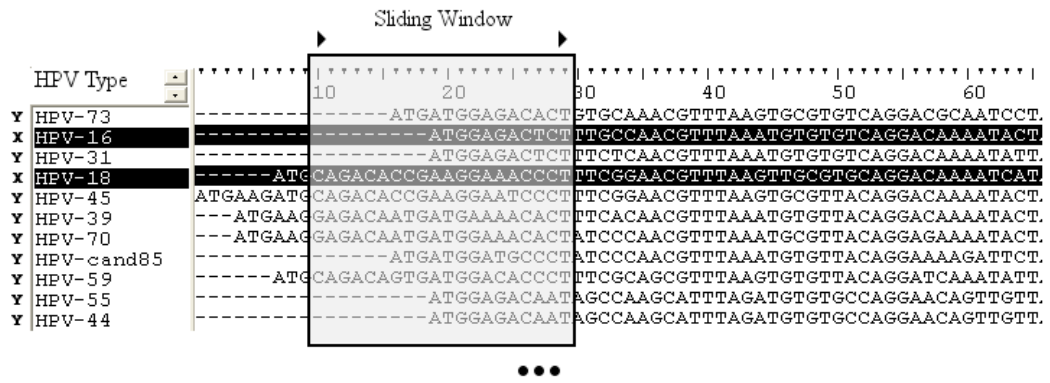


Figure 7–3: A sliding window of a fixed width was used to scan all each HPV gene separately. The sequences in black belong to the set X (carcinogenic HPVs; in this example HPVs 16 and 18), all other sequences belong to the set Y (non-carcinogenic HPVs). The organism is indicated in the column on the extreme left.

from the distinct sets X and Y. The mean of the squared distances computed among the sequence fragments of the carcinogenic taxa from the set X, and denoted here $V(X)$, is computed as follows:

$$V(X) = \frac{1}{(|X| \cdot (|X| - 1)/2)} \sum_{\{x_1, x_2 \in X | x_1 \neq x_2\}} dist_h^2(x_1, x_2), \tag{7.1}$$

and the mean of the squared distances computed only between the sequence fragments from the distinct sets X and Y, and denoted here as $D(X, Y)$, is computed as follows:

$$D(X, Y) = \frac{1}{|X| \cdot |Y|} \sum_{\{x \in X, y \in Y\}} dist_h^2(x, y), \tag{7.2}$$

where $dist_h(x_1, x_2)$ is the Hamming distance between the sequence fragments corresponding to the taxa x_1 to x_2.

Then, the hit region identification function Q is defined as follows:

$$Q = ln(1 + D(X, Y) - V(X)). \tag{7.3}$$

The larger the value of this function for a certain genomic region, the more distinct are the carcinogenic taxa from the non-carcinogenic ones. The use

of the Hamming distance instead of the well-adapted sequence to distance transformations such as the Jukes-Cantor (1969), Kimura 2-parameter (1980) or Tamura-Nei (1993) corrections, is justified by the two following facts: first, often the latter transformation formulae are not applicable to short sequences (remember that in our experiments the sequence lengths, equal to the sliding window width, varied from 3 to 20 nucleotides), and second, most of the well-known transformation models either ignore gaps or assign a certain penalty to them. As the carcinogenicity of HPVs can be related to an insertion or deletion of a group of nucleotides, the gaps should not be ignored but rather considered as valid characters, with the same weight as the other nucleotides, when computing the pairwise distances between the genomic regions.

The time complexity of this algorithm executed with overlapping sliding windows of a fixed width, and advancing one alignment site by step, is $O(l \times n^2 \times w)$, where l is the length of the multiple sequence alignment, n the number of taxa, and w the window width. However, this complexity can be reduced to $O(n^2 \times l)$ if we avoid recomputing the Hamming distance for neighbouring overlapping windows. This can be done by only removing the value of the left column of the sliding window while taking into account the value of added column in the Hamming distance of the sliding window. For a non-overlapping sliding window, the time complexity is $O(n^2 \times l)$. If the width of the sliding window varies, as it was the case in our experiments, the time complexity should be obviously multiplied by the difference between the maximum and minimum window widths. The detailed algorithmic scheme is presented below.

To identifiy a region as a hit, one might use a measure to determine whether the given region has a value of Q higher than a given threshold. However, it is unclear what will be the best value of threshold, since the

Algorithm 3 Algorithmic scheme(MSA, *MSA_L*,*X*, *N(X)*, *Y*, *N(Y)*, *WIN_MIN*, *WIN_MAX*, *S*, *TH*)

Require: MSA: Multiple sequence alignment (considered as a matrix),
MSA_L : Length of MSA,
X: Set of carcinogenic taxa,
N(X): Cardinality of the set X,
Y: Set of non-carcinogenic taxa,
N(Y): Cardinality of the set Y,
WIN_MIN: Minimum sliding window width,
WIN_MAX: Maximum sliding window width,
S: Sliding window step,
TH: Minimum Q value for Hit (i.e., hit threshold).

Ensure: Set of Hit Regions: (win_width, idx, Q), where
win_width : Current sliding window width,
idx : Hit Index (i.e., its genomic position),
Q : Value of the hit region identification function.

for win_width **from** *WIN_MIN* **to** *WIN_MAX* **do**
  for idx **from** 0 **to** $MSA\_L - win\_width$ **with step** S **do**
    MSA_X $\leftarrow$ MSA$[X][idx..idx + win_width]$
    MSA_Y $\leftarrow$ MSA$[Y][idx..idx + win_width]$
    $V(X) \leftarrow D(X,Y) \leftarrow 0$
    for all distinct $i, j \in X$ **do**
      $V(X) \leftarrow V(X) + dist_h^2$(MSA_X[$i$], MSA_X[$j$])
    end for
    $V(X) \leftarrow 2 \times V(X)/(N(X) \times (N(X) - 1))$
    for each $i \in X$ **and** $j \in Y$ **do**
      $D(X,Y) \leftarrow D(X,Y) + dist_h^2$(MSA_X[$i$], MSA_Y[$j$])
    end for
    $D(X,Y) \leftarrow D(X,Y)/(N(X) \times N(Y))$
    $Q \leftarrow ln(1 + D(X,Y) - V(X))$
    if $Q > TH$ **then**
      identify the current region (win_width, idx, Q) *as a hit region*
    end if
  end for
end for

distribution of values of Q might be different in function of the alignment. One possibility could be to rank the Q values and choose a set of highest ones. Moreover, an approach involving the computation of p-values could be implemented to determine the regions that have a value of Q that is different

from the normal Q values of the alignment. Here, we used these two approaches to choose the relevant regions according to their value of Q. To compute the p-value for each given region W_i with a Q value Q_i, Monte Carlo sampling was perfomed, to estimate the distribution of the Q values for a subset of W randomly chosen columns. One million samples were generated and their Q values computed. The p-value of Q_i is then the fraction of samples that obtain a Q value larger or equal to Q_i. It is worth noting that one would expect most of the region with value of Q to have a p-value above 0.001.

7.6 Results, discussion and conclusion

The procedure for identifying hit regions in the 83 available HPV genomes was carried out twice: first, with overlapping windows of width w ($w = 3$ to 20), advancing one alignment site by step, and second, with non-overlapping windows of width 20. The 8 most important HPV genes (see Table 7–3) were scanned in such a way. The scan based on the overlapping windows provided over 35,000 values of Q larger than 0.25. From the best 100 results obtained for each gene, we manually selected (see Table 7–3) the longest contiguous regions (up to 20 nucleotides) corresponding to the largest values of the hit region identification function Q. The values of Q were dependent on the window width, with better results usually associated with small windows.

Table 7–3: Selected high-scoring regions with respect to the values of the hit region identification function Q. The best results for the contiguous regions of size 13 to 20 are reported. The best entry by HPV type (High-Risk, Squam, Adeno) and by gene is presented. The largest values of Q are in bold.

Dataset	*Gene*	Q	*Index*	*Window width*	$D(X,Y)$	$V(X)$.
High-Risk	E1	0.417	695	16	0.74	0.22
Squam	E1	0.345	575	14	0.50	0.08
Adeno	E1	0.353	307	20	0.52	0.09
High-Risk	**E2**	**0.553**	**1289**	**13**	**0.76**	**0.02**
Squam	E2	0.385	613	16	0.47	0.00
Adeno	E2	0.415	1265	20	0.66	0.14
High-Risk	E4	0.480	606	17	0.62	0.00
Squam	E4	0.373	1035	15	0.46	0.01
Adeno	E4	0.395	549	15	0.49	0.00
High-Risk	E5	0.339	88	13	0.41	0.01
Squam	E5	0.401	72	16	0.50	0.00
Adeno	E5	0.363	72	16	0.44	0.00
High-Risk	E6	0.496	725	17	0.69	0.05
Squam	**E6**	**0.531**	**725**	**17**	**0.76**	**0.06**
Adeno	**E6**	**0.521**	**725**	**17**	**0.75**	**0.06**
High-Risk	E7	0.258	206	13	0.34	0.05
Squam	E7	0.263	445	16	0.38	0.08
Adeno	E7	0.262	110	16	0.40	0.10
High-Risk	**L1**	**0.574**	**241**	**14**	**0.79**	**0.02**
Squam	L1	0.294	1159	15	0.34	0.00
Adeno	L1	0.302	1181	17	0.56	0.20
High-Risk	L2	0.310	1751	14	0.65	0.28
Squam	L2	0.320	1916	15	0.38	0.00
Adeno	L2	0.313	1914	17	0.37	0.00

For instance (see Table 7–3), for larger window sizes, the largest values of Q were found during the scans of genes E2 and E6 for all types of HPVs, with the exception of the overall best score obtained during the scan of the gene L1 for the High-Risk HPV types (the value of 0.574 for a 14-nucleotide region starting with the index 241, see Table 7–3). For windows of small width, the largest values of Q were observed during the scan of the gene E4 for the High-Risk HPV category but in Table 7–3 we show only the best results for the longer contiguous regions of size 13 to 20 nucleotides. All the regions presented in Table 7–3 have a p-value at most 10^{-6}.

Figure 7–4 depicts the progressive results obtained during the scan of the L1 gene and the High-Risk HPVs (HPVs-16 and 18) with the non-overlapping windows of size 20 nucleotides. The highest score, for the non-overlapping windows of size 20 among all genes and all types of HPV-caused cancers, of the Q function ($Q = 0.55$) was obtained for this gene.

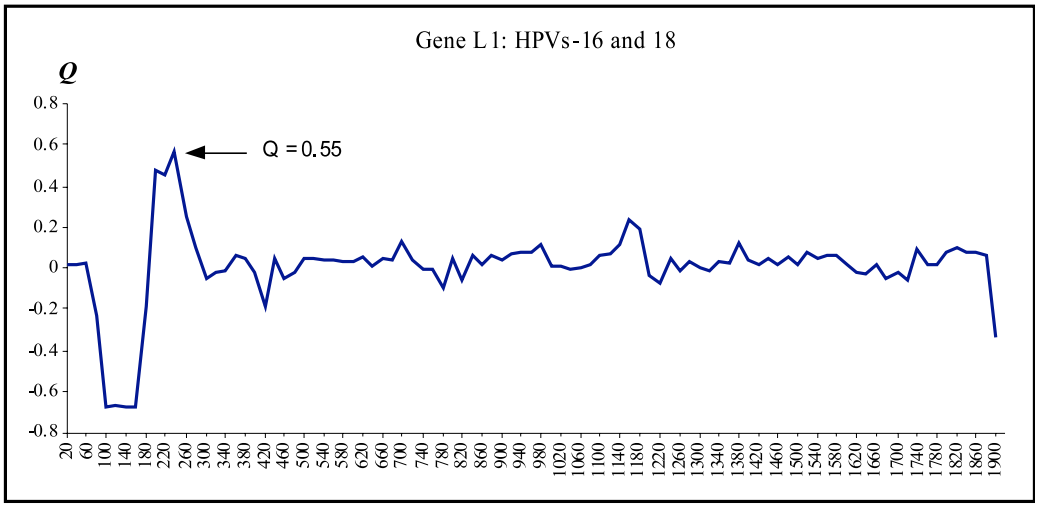


Figure 7–4: The variation of the hit identification function Q for the High-Risk HPVs (HPVs-16 and 18) obtained with the non-overlapping sliding widows of width 20 during the scan of the L1 gene. The abscissa axis represents the window position.

As most of the largest values of Q were obtained for the genes E2 and E6, we also present in Figure 7–7 and 7–8 the progressive results diagrams illustrating the scan of these genes with the non-overlapping windows of size 20. The largest values of the hit region identification function Q are usually found during the scan of the genes E2 and E6. Moreover, we found that in these two genes the number of regions obtaining p-values less than 0.001 is the largest. For instance, in gene E6, three large regions of size between 40 nucleotides and 60 nucleotides have a p-value less than 0.001 (Figure 7–5 and 7–6). The last region of figure of E6 surprisingly corresponds to a PDZ domain-binding motif (-X-T-X-V) at the carboxy terminus of the protein, which is essential for targeting PDZ proteins for proteasomal degradation. Such proteins include hDlg, hScrib, MAGI-1, MAGI-2, MAGI-3, and MUPP1 [36]. The interaction between the E6 protein and hDLG or other PDZ domain-containing proteins could be an underlying mechanism in the development of HPV-associated cancers [239] .

It is worth noting that according to recent findings the high expression of E6 and disruption of E2 might play an important role in the development of HPV-induced cervical cancer [247]. As result of E6 high expression, the immune system is potentially evaded [186]. Disruption of the gene E2 was observed in invasive carcinomas [31] and in high-grade lesions [99]. Surprisingly, the overall largest value of Q was obtained for a specific region of the L1 gene. This underlines the possible use of our method for investigating particular regions of capsidal proteins in relation with vaccine design. It has been shown that linear epitopes within the protein L1 that induce neutralizing antibodies exist [39].

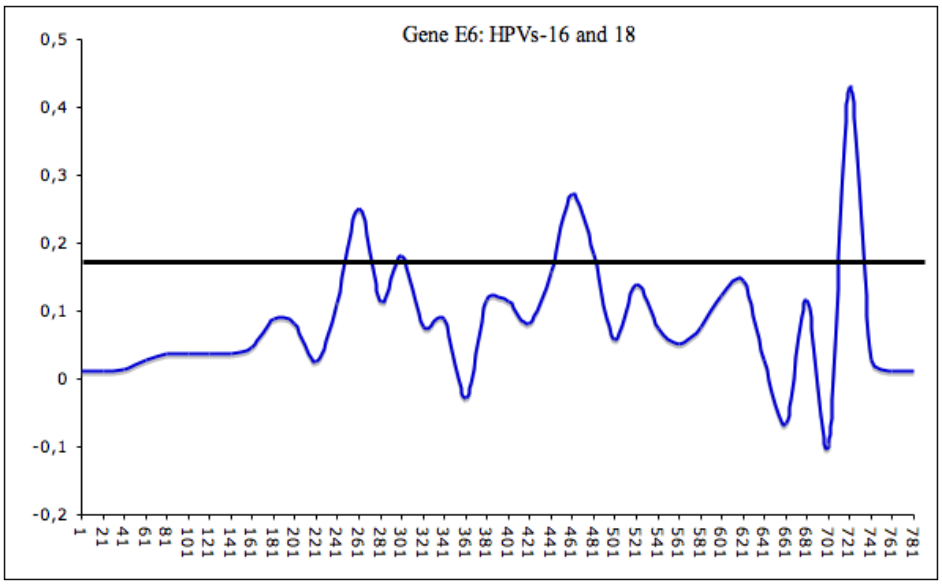


Figure 7–5: The variation of the hit identification function Q for the High-Risk HPVs (HPVs-16 and 18) obtained with the non-overlapping sliding widows of width 20 during the scan of the E6 gene. The horizontal line cutting the graph represents the threshold of p-value less than 0.001. The abscissa axis represents the window position.

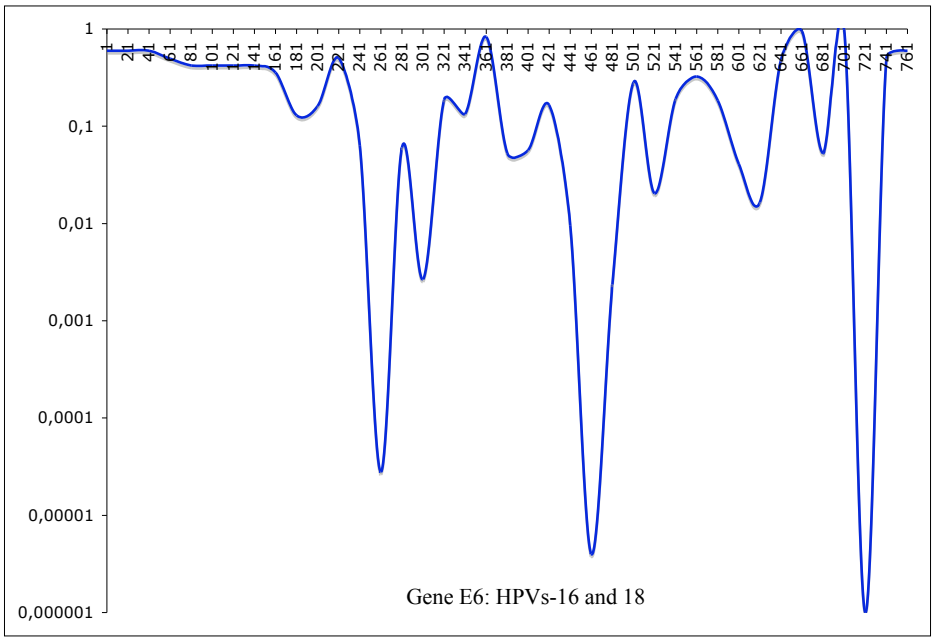


Figure 7–6: The variation of the p-value in the different region of the alignment for the High-Risk HPVs (HPVs-16 and 18) obtained with the non-overlapping sliding widows of width 20 during the scan of the E6 gene. The abscissa axis represents the window position.

We observed that the results obtained depend on the window width. As substitutions affect individual sites whereas indels often involve several consecutive nucleotides, small window sizes will tend to favor the former. However, the use of the Hamming distance, which does not ignore gaps in calculation, and variable window width allows us to account for both substitution and indel events. In the future, it would be interesting to study in more detail, in collaboration with virologists, all genomic regions providing the highest scores of the hit region identification function Q (particular attention should be paid to the E2, E6 and L1 genes), and to determine, for each selected region, the evolutionary events (substitutions or indels) responsible for the observed differences in the carcinogenic and non-carcinogenic HPVs, and then establish at which level (i.e. on which branch) of the associated gene phylogeny this event has occurred. It may also be interesting to consider merging our results to those given by methods for detecting sequences under lineage-specific selection such as DLESS [216]. Next, we plan to compare this work with other approaches on the computational virology, which used some simpler methods, such as signatures, to analyze other viruses. Another interesting development would be to design more sophisticated statistical tests allowing one to measure the statistical significance of the obtained results.

Acknowledgement. B.D. is an NSERC fellow. We thank Alix Boc and Emmanuel Mongin for their useful comments.

Additionnal Materials. Additionnal materials related to this study is available at <http://ancestors.bioinfo.uqam.ca/articles/JCB2009/supplemental.zip>. These materials contain the data used and the whole results for all scanned genes with different window width.

Appendix

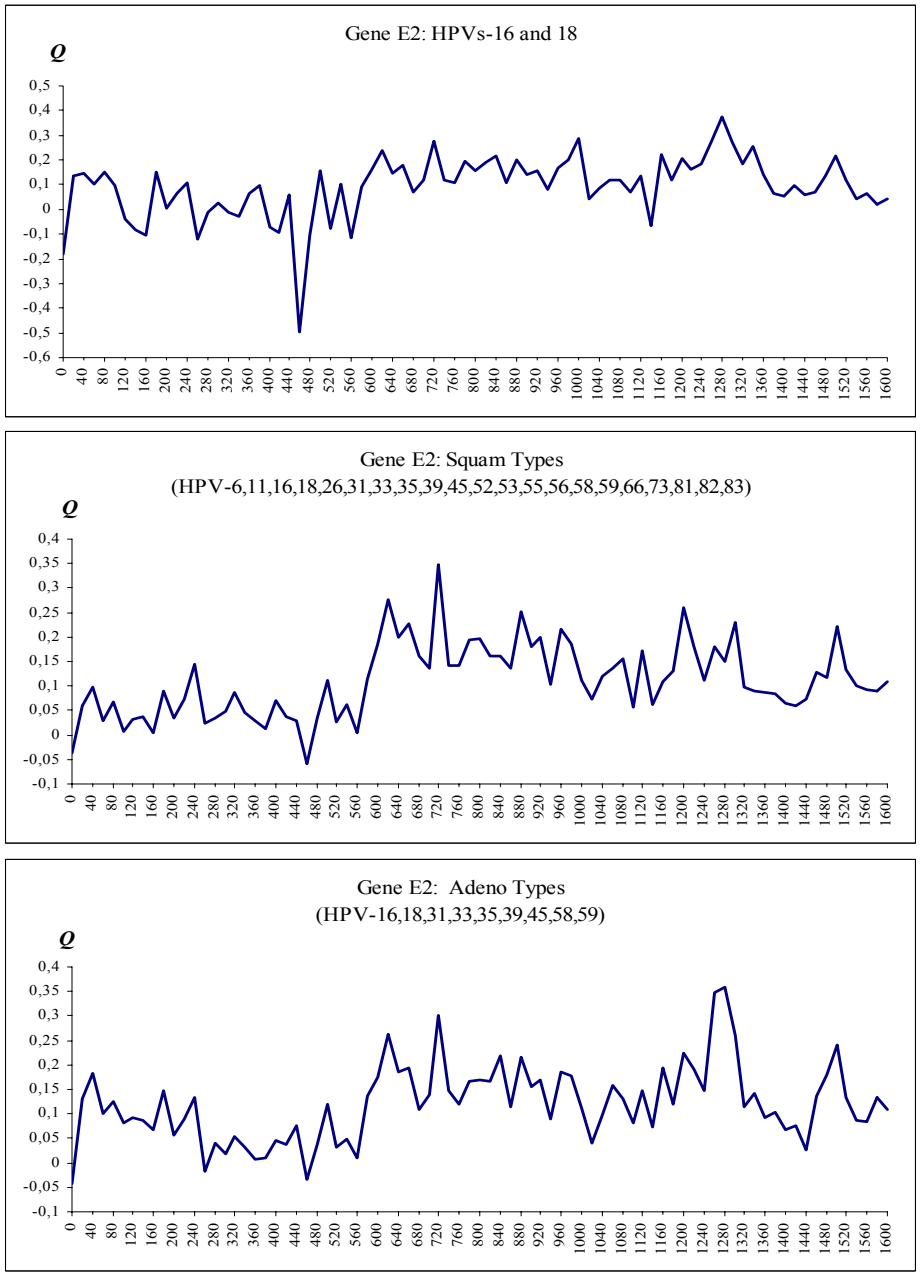


Figure 7–7: The variation of the hit identification function Q for: (a) High-Risk HPVs (HPV-16 and 18), (b) Squam cancer causing HPVs, and c) Adeno cancer causing HPVs obtained with the non-overlapping sliding widows of width 20 during the gene E2 scan.

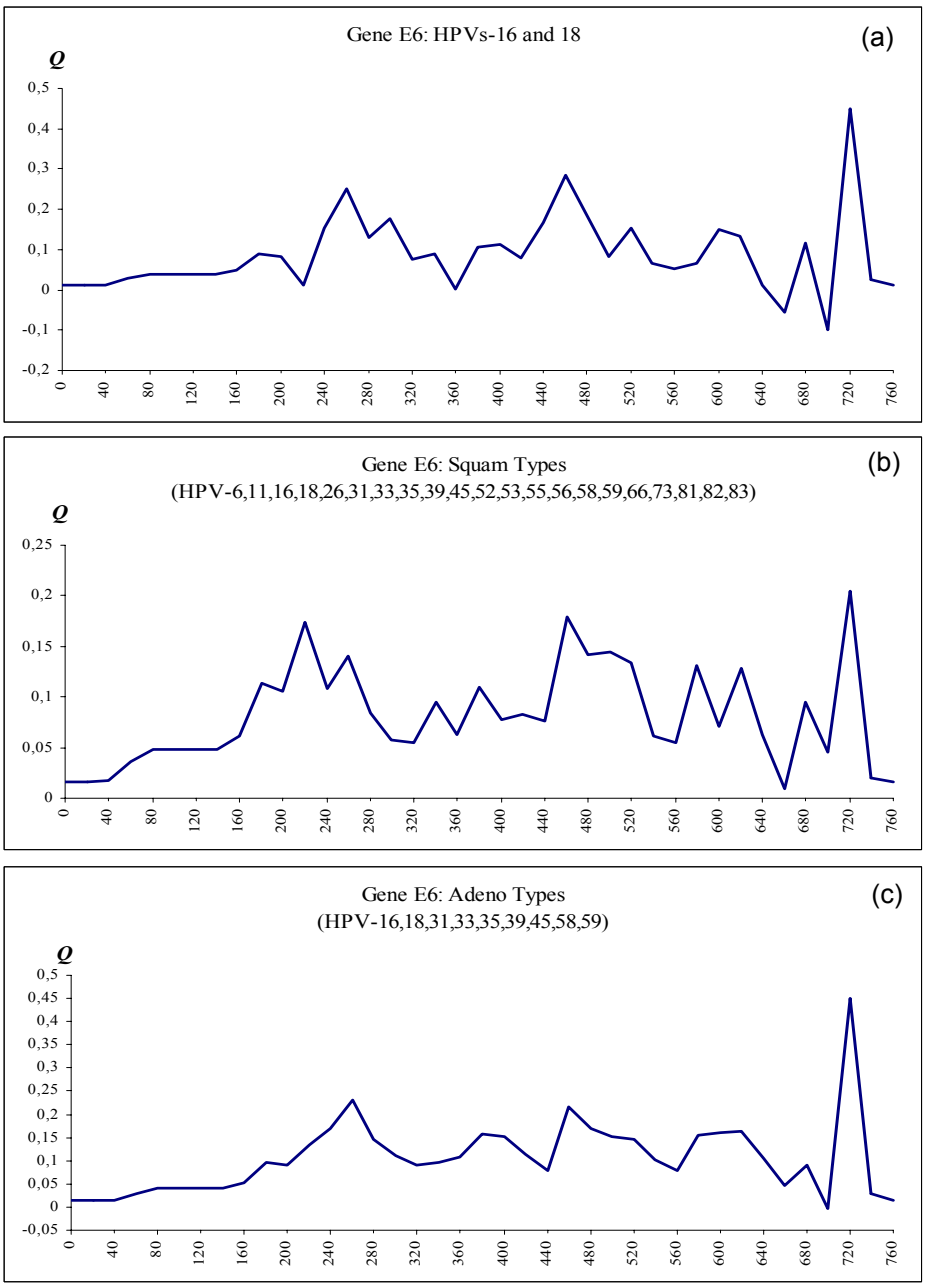


Figure 7–8: The variation of the hit identification function Q for: (a) High-Risk HPVs (HPV-16 and 18), (b) Squam cancer causing HPVs, and c) Adeno cancer causing HPVs obtained with the non-overlapping sliding widows of width 20 during the gene E6 scan.

CHAPTER 8
Conclusion

In this thesis, we proposed in Chapter 3 an exact algorithm for the problem of reconstructing the most likely scenario of insertions and deletions capable of explaining the gaps observed in a given alignment according to a given phylogenetic tree [58, 62]. Furthermore, we also designed a new probabilistic framework for indel analyses. The new probabilistic framework provides a way of weighing insertions and deletions of various lengths against each other. It also provides an accurate probabilistic model of indels, an exact and heuristic algorithm for the reconstruction of indel scenarios, and allows the estimation of the uncertainty for each part of the solution. Similarly to the statistical alignment approaches, which unfortunately remain too slow for genome-wide reconstructions, our method seeks to gain a richer insight into ancestral sequences and evolutionary processes of more than 20 taxa. This framework is the core of the Ancestors 1.0 program available at: <http://ancestors.bioinfo.uqam.ca/ancestorWeb> and presented in chapter 4 of this thesis. It will be integrated soon into the pipeline of the project of the ancestral mammalian reconstruction initiated by David Haussler from the University of California at Santa Cruz (UCSC), with the collaboration of several other universities such as Pennsylvania State University and McGill University. Apart of ancestral sequence reconstruction, we showed the utility of the indel model to (1) improve multiple sequence alignment and phylogenetic tree reconstructions, (2) and to replace the topological distance of Robinson and Foulds [201] or multiple sequence alignment scores in simulation procedures (chapter 5). Furthermore, we presented two applications of our framework

in the studies of phages (chapter 6) and Human Papilloma Viruses (chapter 7). In the phages analyses, we proposed a new approach to reconstruct the phage classification and an extension of the ancestral sequence reconstruction framework to allow horizontal gene transfers and partial genomic data. In chapter 6, our probabilistic framework has been used to analyze the relation of indel evolutionary events and the phylogenetic aspect of carcinogenicity [55] . We also provided an efficient way for identifying specific carcinogenic regions according to their evolutionary events, and proposed the first whole genome classification of the Human Papilloma Viruses family [54].

8.1 Major contributions

The most important contributions of my thesis can be summarized as follows:

- Providing algorithms to solve the indel maximum likelihood problem for large data sets.
- Providing web server and visualization tools for the inference of ancestral sequences and their uncertainties.
- Providing databank of phages ancestral sequences.
- Providing a method to assess the joint inference of phylogenetic tree and multiple sequence alignment, and using it to select the best tree-alignment pair.
- Proposing an algorithm to identify specific regions according to the relation between their carcinogenic evolution and the mutation events.
- Providing framework for ancestral sequence reconstruction for reticulate evolutionary genomes allowing horizontal gene transfer events.

8.2 Perspectives

In the future, we intend to correct certain types of small-scale multiple sequence alignment errors using our ability to reconstruct ancestral sequences

and their uncertainty. Given an original imperfect alignment, we will predict ancestral profile sequences based on our framework. The algorithm will realign each extant sequence to its most recent ancestor using our pairwise profile-profile alignment (see Appendix D). Then recent ancestors will be realigned with their ancestor. This approach is similar to the MAVID method [25]. However, MAVID does not use profiles, and considers gaps as fifth symbol, while the introduction of affine gap penalties as described in appendix D would be preferable.

We also propose to use our indel likelihood score presented in chapter 5, combined with the substitutions likelihood score to design a new method (extending our phylogenetic-Hidden Markov Model approach [58]) for the joint inference of phylogenies and multiple sequence alignment [108] that can be applicable to large genomic data. For this purpose, we will design an iterated algorithm that can build partially accurate multiple sequence alignment from unaligned sequences using one of existing alignment methods, then infer ancestral sequences using the developed algorithm. Finally, the uncertainties related to the ancestral prediction can be used to refine the alignment and re-estimate the ancestral characters, followed by the tree topology rearrangement using the existing techniques such as Nearest Neighbor Interchange (NNI), Subtree Pruning and Regrafting (SPR) and others. Branch lenghts estimations can be handled apart as individual branch length optimization. Variant of the numerical optimization of branch lenghts showed in PHYML [101] would be adequate for this problem. This scenario could be iterated until a convergence is obtained. If the tree rearrangement steps improve highly the agreement in each iteration, the method will converge quickly, and will be preferable to the available profile-HMM based one SATCHMO [72] that do alignment and tree reconstruction in parallel. Another important task would be refining

the visualization of the ancestral sequence reconstruction such that very large datasets could be efficiently analyze. Finally, once all the mentioned tools are operational, we will have an accurate and easy way to use indels likelihood framework that can play an important role in genome analyses, studying for example the impact of indels on gene regulation, functional region annotation and more.

APPENDIX A
Computational Reconsruction of Ancestral DNA Sequences

A.1 Preface

This appendix contains a presentation of the motivation of ancestral sequences reconstruction, the different steps and their challenges. It is published as a book chapter [15].

A.2 Abstract

This chapter introduces the problem of ancestral sequence reconstruction: given a set of extant orthologous DNA genomic sequences (or even whole genomes), together with a phylogenetic tree relating these sequences, predict the DNA sequence of all ancestral species in the tree. Blanchette et al. (2004) have shown that for certain sets of species (in particular, for eutherian mammals), very accurate reconstruction can be obtained. We explain the main steps involved in this process, including multiple sequence alignment, insertion and deletion inference, substitution inference, and gene arrangement inference. We also describe a simulation-based procedure to assess the accuracy of the reconstructed sequences. The whole reconstruction process is illustrated using a set of mammalian sequences from the CFTR region.

A.3 Introduction

Following the completion of the human genome sequence, there is now considerable interest in obtaining a more comprehensive understanding of its evolution [122, 233, 197]. Patterns of evolutionary conservation are used to screen human DNA mutations to predict those that will be deleterious to protein function and to identify noncoding sequences that are under negative selection, and hence may perform regulatory or structural functions [161, 41,

10]. Long periods of conservation followed by sudden change may provide clues to the evolution of new human traits [98, 78]. All of these efforts depend, directly or indirectly, on reconstructing the evolutionary history of the bases in the human genome, and hence on reconstructing the genomes of our distant ancestors.

Although some information about ancestral species has been irrevocably lost during evolution, there is still the possibility that large regions of the genomes of ancestral species with many modern descendants can be approximately inferred from the genomes of modern species using a model of molecular evolution. Indeed, it has recently been reported that in the specific case of mammalian evolution, ancestral genome reconstruction was possible to a surprising degree of accuracy [16].

The ideal target species for a genomic reconstruction is one that has generated a large number of independent, successful descendant lineages through a rapid series of early speciation events. In this case, the problem can be viewed as attempting to reconstruct an original from many independent noisy copies. In the limit of an instantaneous radiation, the accuracy of the reconstruction approaches 100% exponentially fast as the number of copies increases. From the Cretaceous period, a good choice for reconstruction would be the genome of the eutherian ancestor, as this species is believed to have spawned the relatively rapid radiation of the different lineages of modern placental mammals [76, 225]. This ancient species also has the added advantage of being a human ancestor, so its reconstruction, however speculative, may shed additional light on our own evolution, perhaps helping to explain features of the human and other modern mammalian genomes.

In this chapter, we describe the set of computational approaches and tools that exist for reconstructing ancestral sequences and for estimating the

accuracy of such a reconstruction. This area being relatively new, there is not a single tool that performs all the steps involved in the reconstruction. Instead, tools developed by different authors need to be used sequentially. The methods are illustrated on a 1.8Mb region of mammalian genomes, containing the CFTR gene, sequenced by the ENCODE project [234]. Much of this chapter is derived from Blanchette et al. (2004).

A.4 Materials

A.4.1 Sequence data

To reconstruct ancestral sequences, orthologous DNA regions from as many descendants as possible need to be compared. The more orthologous sequences are available, the more accurate the reconstruction will be, provided accurate evolutionary models are used. For vertebrate sequences, a good repository of complete genome sequences is the UCSC Genome Browser (*http://genome.ucsc.edu*, [128]). Besides raw DNA sequences, multiple genome alignments, and various types of genome annotation are accessible from the same site.

For the purpose of this chapter, we illustrate the process of ancestral sequence reconstruction using a 1.8Mb region of the human genome including the CFTR gene, together with orthologous regions from 19 other mammals (available at the UCSC Genome Browser). This deep coverage is not currently available over all the genome, but only for the targeted sequencing of the ENCODE project [234].

A.4.2 Phylogenetic information

An important component of ancestral sequence reconstruction is the knowledge of the phylogenetic relationships among the species being compared. Knowing the correct tree topology and estimating the length of its branches is crucial for an accurate reconstruction, as well as for estimating the accuracy

of that reconstruction through simulations. For many sets of species, accepted phylogenetic trees are now available (see for example [153], and [76]). For others, the exact phylogenetic relationships remain unclear and need to be inferred prior to reconstruction, using programs like Phylip [85], PhyML [101], or MrBayes [118]. These tools are also necessary to estimate the branch lengths of the phylogenetic tree using a maximum likelihood approach.

A.4.3 Sequence annotation

In some cases, functional annotation of extant sequences can be used to obtain more accurate reconstruction of ancestral sequences. This is particularly the case for coding region annotation and repetitive region annotation. For metazoans, a good source of such annotations is the UCSC genome browser and the Ensembl Genome Browser (*http://www.ensembl.org*).

A.5 Methods

This section introduces the techniques that have been developed for predicting ancestral DNA sequences based on their extant descendants, and for estimating the accuracy of the reconstruction. We illustrate this reconstruction process and the type of information that can be derived from it using 1.8 Mb region surrounding the *CFTR* gene in mammals (see [16] for more details).

A.5.1 Predicting ancestral sequences

The prediction of ancestral genomes can be decomposed into four main steps. A crucial first step toward the reconstruction is to build an accurate multiple alignment of the extant orthologous sequences, thus establishing orthology relationships among the nucleotides of each sequence. Second, the process of indel reconstruction determines the most likely scenario of insertions and deletions that may have led to the extant sequences. Third, substitution history is reconstructed using a maximum likelihood approach. The

last step involves dealing with genome rearrangements (inversions, transpositions, translocations, duplications, and chromosome fusions, fissions, and duplications).

Multiple sequence alignment. Given a set of orthologous sequences, the multiple alignment problem consists of identifying (by aligning them together) the sets of nucleotides derived from a common ancestor through direct inheritance or through substitution. Many approaches have been developed to align multiple, large genomic regions. Some of the most popular approaches include programs like MAVID [25], MLAGAN [27, 42], and TBA [17]. All these approaches fall under the category of progressive alignment methods, and require the prior knowledge of the topology of the phylogenetic tree that relates the extant sequences compared (see Section A.2). The threaded blocks aligner (TBA) program, based on the well-established pair-wise alignment program BLASTZ [212], has been shown to be particularly accurate for aligning mammalian sequences and is thus a tool of choice for ancestral reconstruction for these species. The program is available at: *http://www.bx.psu.edu/miller_lab/.* The multiple sequence alignment problem is discussed in more details in Millers chapter in this book.

Indel Reconstructing. Given a multiple sequence alignment of the repeat-soft-masked extant sequences and a phylogenetic tree with known topology and branch lengths, the next step consists of predicting, for each ancestral node in the tree, which columns of the alignment correspond to ancestral bases, and which correspond to nucleotides inserted after the ancestor. While the problem of parsimonious indel inference has recently been shown to be NP-Hard [35], good heuristics have been developed by Fredslund et al. [91], Blanchette et al. [16], and Chindelevitch et al. [35]. Currently, the only publicly available program for indel reconstruction is the inferAncestors program

based on the greedy approach of Blanchette et al. [16]. This section describes briefly how the program works.

Given a multiple alignment, all the gaps in the alignment are first marked as unexplained. The algorithm iteratively selects the insertion or deletion, performed along a specific edge of the tree and spanning one or more columns of the alignment, that yields the largest number of alignment gaps explained per unit of cost. The number of gaps explained by a deletion is the number of unexplained gaps in the subtree above which the deletion occurs. The number of gaps explained by an insertion is the number of unexplained gaps in the complement of the subtree above which the insertion occurs. The costs can be defined heuristically. The cost of a deletion is given by 1 + 0.01 log(L) 0.01 b where L is the length of the deletion and b is the length of the branch along which the event takes place. The cost of an insertion is given by 1 + 0.01 log(L) 0.01 b r, where L and b are defined as above and r is a term that takes value 0.5 if the repetitive content of the segment inserted is more than 90%. Once the best insertion or deletion has been identified, its gaps are marked as explained. This does not preclude them from being part of other indels, but they will not count in their evaluation. Finally, heuristics are used to reduce errors due to incorrect alignment, in particular to reduce the problems caused by two repetitive regions from two distantly related species mistakenly aligned to each other, with other species having gaps in that region.

Substitutions reconstruction. After having established which positions of the multiple alignment correspond to bases in the ancestor, the inferAncestors program predicts which nucleotide (A, C, G, or T) was present at each position in the ancestor using the standard posterior probability approach [259] based on a dinucleotide substitution model where substitutions at two adjacent positions are independent except for CpG, whose substitution rate

to TpG is ten times higher than those of other transitions [215]. This phase of the reconstruction relies on the availability accurate branch length estimates for the phylogenetic tree, which can be obtained as described in Section A.2.

The inferAncestors program. The inferAncestor program, available from *http://www.mcb.mcgill.ca/~blanchem/software*, integrates the steps of indel and substitution inference. The algorithm takes as input a multiple alignment in fasta format, together with a phylogenetic tree in New Hampshire format. The program outputs a predicted ancestral sequence for each internal node of the phylogenetic tree. Two other files are output, describing the confidence of the prediction made for each base of each ancestral sequence. The first describes the confidence in the prediction of presence or absence of a base at each position of each ancestral sequence. The second describes the confidence of the actual nucleotide (A, C, G, or T) predicted. The inferAncestor program is written in C++ and has been tested on Linux and Mac OS X.

Genome Rearrangements. To complete the inference of ancestral genomes, the ancestral DNA sequences inferred for each block of orthologous sequences need to be ordered into a single, contiguous genome. This problem is made challenging by the presence of genome rearrangements (inversions, transpositions, translocations, and duplications/losses). One of the most popular computer programs for inferring ancestral gene arrangement is MGR ([21], *http://www.cse.ucsd.edu/groups/bioinformatics/MGR*), which is described in details in Bourques chapter in this book.

A.5.2 Assessing reconstruction accuracy through simulations

This section describes a simulation-based method for assessing the accuracy of the reconstructed ancestor. An alternate approach based on retrotransposons is described in [16].

To assess the reconstructability of ancestral genomic sequences from their extant descendants, the simplest method is to use simulations of sequence evolution. Starting from a known (but synthetic) ancestral sequence, we let the sequence evolve along the branches of the tree, until the leaves are reached. The ancestral sequence reconstruction procedure is then applied to the set of simulated leaves, and the prediction made is compared to the known ancestral sequence.

The simulation program Simali (*http://www.bx.psu.edu/miller_lab/*), based on the Rose program [226], can be used to mimic the evolution of sequences under no selective pressure. Given a phylogenetic tree, the program simulates sequence evolution by performing random substitutions, deletions, and insertions along each branch, in proportion to its length. The program allows for the insertion of retrotransposons, which is an important source of error in sequence alignment, and thus in ancestral sequence reconstruction.

To assess the reconstructability of ancestral mammalian genomic sequences, Blanchette et al. [16] performed a series of computational simulations of the neutral evolution of a hypothetical 50Kb ancestral genomic region into orthologous regions in 20 modern mammals (Figure A–1). The simulations are based on the phylogenetic tree inferred by Eizirik et al. [76] on a set of genes for a large set of mammals. Substitutions follow a context-independent HKY model [105] with Ts/Tv = 2, p(a) = p(t) = 0.3, and p(c) = p(g) = 0.2, except that substitution rates of CpG pairs are ten times higher than other rates [215]. Deletions are initiated at a rate of about 0.056 times the substitution rate, their length is chosen according to a previously reported empirical distribution [130] that ranges between one and 5000 nucleotides, and their starting

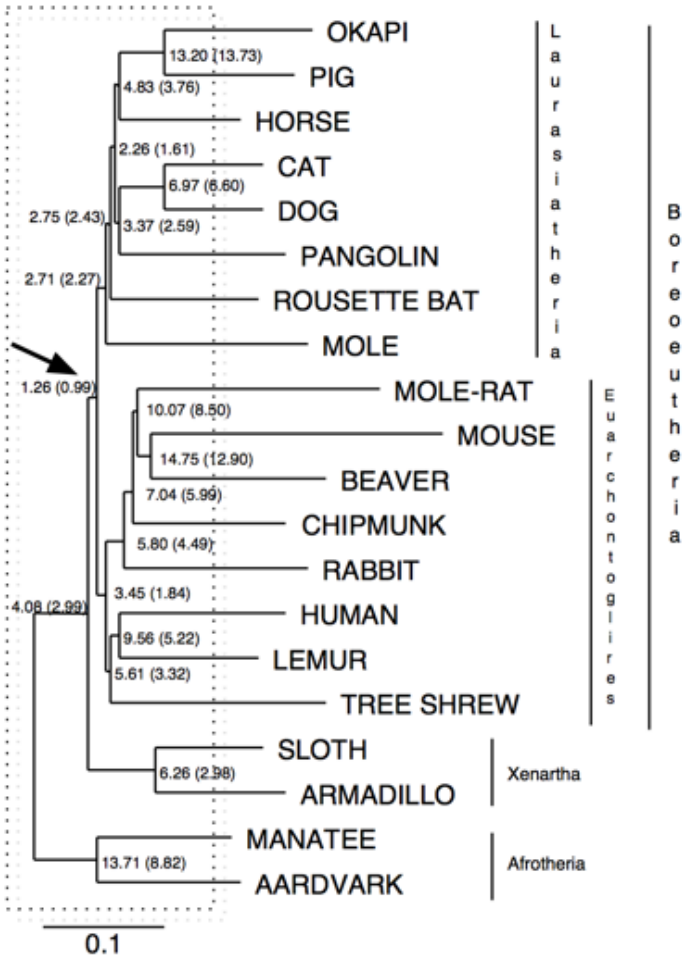


Figure A–1: Estimated reconstructability of ancestral mammalian sequences. Average base-by-base error rate in the reconstruction of each simulated ancestral sequence. The error rate shown is the sum of the percentages of bases that are missing, added, or mismatched as a result of errors in the reconstruction, averaged over one hundred simulations of sets of orthologous sequences of length approximately 50kb. Error rates are given first for all regions, and in parentheses for non-repetitive regions only. The species names at the leaves only indicate what organisms we simulated; no actual biological sequences were used here. The tree topology and branch lengths are derived directly from Eizirik et al. [76].

point is uniformly distributed. Insertions occur randomly according to a mixture model. Small insertions (of size between 1 and 20nt) occur at half the rate of deletions, their size distribution is empirically determined [130] and their content is a random sequence where each nucleotide is chosen independently from the background distribution. They also simulate the insertion of retrotransposons. For this they used a library of 15 different types of transposable elements chosen to cover the large majority of repetitive elements observed in well-studied mammals [127]. The rate of insertion of each repeat varies from branch to branch, so that certain retrotransposons (like ALUs, SINEs B2, BOV) are lineage-specific, while others (L1, LTR, DNA) are both present in the sequence at the root of the tree (with a range of decaying level) and can be inserted along any branch. The code and parameters used for our simulations are available with the Simali package.

After generating a set of simulated sequences, the sequences are first soft-repeat-masked using RepeatMasker [219] and then aligned using one of the methods in A.1.1. The repeat-masked multiple alignment is then fed to the inferAncestors program, which produces a prediction of the ancestral sequence at each internal node of the phylogenetic tree. To compare the actual ancestral sequence generated by simulations to the predicted ancestral sequence, we align them and count the number of missing bases (those present in the actual ancestor but not in the reconstruction), added bases (present in the reconstruction but not in the actual ancestor), and mismatch errors (positions in the reconstruction assigned the incorrect nucleotide). The sum of the rates of all three types of errors, calculated separately at each ancestral node in the phylogenetic tree, is used to estimate the reconstructability of a given ancestor.

In the case of mammalian sequences, Blanchette et al. [16] used the above simulation-based procedure to show that the sequence of certain mammalian

ancestors can be reconstructed remarkably accuractely. Figure A–1 shows that under this phylogenetic tree with a relatively rapid placental mammalian radiation, the neutral non-repetitive regions of the Boreoeutherian ancestral genome that have evolved under their simple model should be reconstructable with about 99% base-by-base accuracy from the genomes of 20 present-day mammals. Repetitive regions are not reconstructed as accurately because they are more often involved in misalignments, which can result in incorrect predictions. Nonetheless, even counting errors in repetitive regions, the total accuracy is more than 98%. The simulations suggest that even in the non-repetitive regions, much of the difficulty of the reconstruction problem lies in the computation of the multiple alignment, as a reconstruction based on the correct multiple alignment derived from the simulation itself (and thus unavailable for actual sequences) had less than half the number of reconstruction errors. Examining reconstructions made using smaller subsets of this set of 20 species, it was found that, including repetitive regions, an accuracy of about 97% can be achieved using only ten species chosen to sample most major mammalian lineages (Figure A–2). Sampling only five of the most slowly evolving lineages yields an accuracy of about 94%. Little is gained with our current reconstruction procedures by adding more than ten species because the risk of misalignment increases, while the unavoidable loss of information in the early branches persists.

A.5.3 Reconstruction of actual mammalian sequences

Blanchette et al. [16] applied the reconstruction method described above to actual high-quality sequence data from a region containing the human

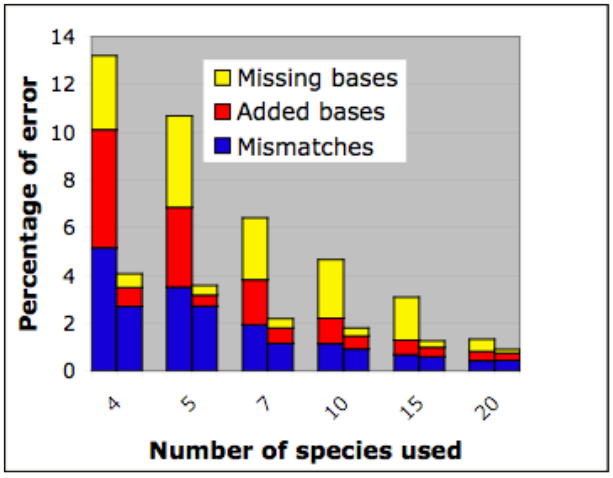


Figure A–2: Estimated reconstructability of the Boreoeutherian ancestor. Fraction of the simulated Boreoeutherian ancestral sequence reconstructed incorrectly as a function of the number of extant species used for the reconstruction. For each number of species used, results are given counting all bases (left columns) and only non-repetitive bases (right columns). Species are added in the following order: human, cat, chipmunk, sloth, manatee, rousette bat, mole, pig, beaver, tree shrew, horse, pangolin, mouse, armadillo, aardvark, okapi, dog, mole-rat, rabbit, lemur.

CFTR locus, using 18 additional orthologous mammalian genomic regions generated by the NISC Comparative Sequencing Program ([234], *www.nisc.nih.gov*). Simulations on synthetic data like those described above indicate that for the topology and set of branch lengths for these 19 species, the ancestral sequence that can be the most accurately reconstructed based on the sequences available is the Boreoeutherian ancestor, and that neutrally evolving regions of this ancestral genome can be reconstructed with an accuracy of about 96%. On a site-specific basis, simulations suggest that more than 90% of the bases of the predicted ancestor can be assigned confidence values greater than 99%. The reconstructed ancestor and site-specific confidence estimates are available at http://genome.ucsc.edu/ancestors.

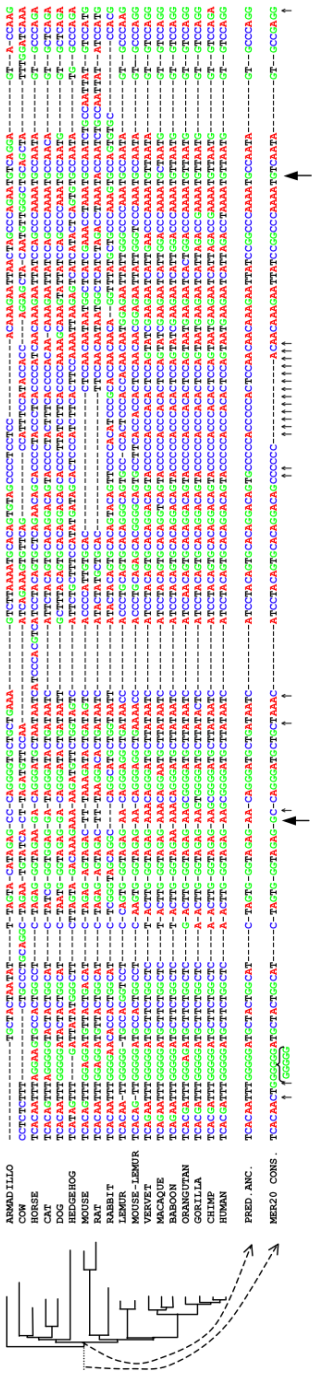


Figure A–3: Example of reconstruction of an ancestral Boreoeutherian sequence based on actual orthologous sequences derived from a MER20 retrotransposon. Arrows indicate positions where the reconstructed ancestor differs from the MER20 consensus. Longer arrows indicate the positions where the knowledge of the MER20 consensus sequence would have changed the ancestral base prediction. The position of the human sequence displayed is chr7:115,739,755-115,739,899 (NCBI build 34). The alignment of the flanking non-repetitive DNA (not shown) verifies that the sequences from the different species are in fact orthologous. The tree and branches are derived directly from [76].

Figure A–3 illustrates the reconstruction in a non-coding region of the CFTR locus that exhibits a typical level of sequence conservation. This region is located in a 32Kb intron of the CAV1 gene, about 13Kb from the 5exon. The bases in this region are relics left over from the insertion of a MER20 transposon sometime prior to the mammalian radiation and are thus unlikely to be under selective pressure. Notice that despite the fact that the alignment of certain species (in particular, mouse, rat, and hedgehog) appears somewhat unreliable, the inference of the presence or absence of a Boreoeutherian ancestral base at a given position is quite straightforward given the alignment, and so is, to a lesser extent, the prediction of the actual ancestral base itself. The MER20 consensus is shown for comparison. Most positions where the reconstructed Boreoeutherian ancestral base disagrees with the MER20 consensus are likely due to substitutions in this MER20 relic that predated the Boreoeutherian ancestor, since the support of the reconstructed base is very strong in the extant species. If the MER20 consensus sequence is used as an outgroup in the reconstruction procedure, only two bases (indicated by a longer arrow) are reconstructed differently, indicating that the reconstructed ancestral sequence is very stable and most of it is likely to be correct.

A.6 Notes

The accuracy of the reconstruction depends crucially on the length of the early branches of the phylogenetic tree. In the context of the ancestral mammalian sequence reconstruction, Blanchette et al. [16] have shown that if the major placental lineages had diverged instantaneously, they would be able to reconstruct the simulated Boreoeutherian ancestral sequence, including repetitive regions, with less than 1% error. In contrast, if the early branch lengths inferred by Eirizik et al. [76] turned out to underestimate the actual

lengths by a factor of two, the error rate would jump to 3%, and to 6% if they were underestimated by a factor of four.

One of the non-intuitive results presented by Blanchette et al. [16] is the observation that more ancient ancestral genomes can often be reconstructed more accurately than their more recent descendants. Why exactly is this so? For simplicity, consider the case of reconstructing a single binary ancestral character state in the root species (e.g. purine vs pyrimidine at a given site) under a simple model in which the prior probability distribution on the ancestral character is uniform, substitution rates are known, symmetric, homogeneous, and not too high, and the total branch length in the phylogenetic tree from the root ancestor to each of the modern species is the same (i.e. assume a molecular clock). Here each of n modern species has a state that differs from the ancestral one with the same probability p ¡ 1/2. If the tree exhibits a star topology, in which each of the modern species derives directly from the ancestor on an independent branch, then it is clear that the maximum likelihood and Bayesian maximum a posteriori reconstructions of the ancestral character agree, and the reconstructed state is the one that is most often observed in the n modern species. The probability of an error in reconstruction is:

$$\sum_{\lceil k=\frac{n}{2} \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}$$

which is at most $[4p(1-p)]^{n/2}$ [113], [142](*Lemma 5 p.479*). This error approaches zero exponentially fast as n increases. The star topology has a kind of phase transition where the ancestor becomes highly reconstructable once enough present day sequences are available to compensate for the length of the branches leading back to the ancestor.

In contrast, a non-star topology such as a binary tree that has the same total root-to-leaf branch length and the same number n of modern species at

the leaves has two nonzero length branches from the root ancestor R leading to intermediate ancestors A and B, and information is irrevocably lost along these two branches. No matter how large the number n of modern descendant species derived from A and B, one can do no better at reconstructing the state at R than if one knew for certain the state in its immediate descendants A and B. Even with this knowledge, the accuracy of reconstruction of R from A and B will be strictly less than 100% for all reasonable models and nonzero branch lengths. The reconstruction gets poorer the longer the branch lengths are to A and B. This extends to the case where the ancestor R being reconstructed has a bounded number of independent immediate descendants and to the case where descendants of an earlier ancestor of R (outgroups) are also available. The long branches connecting them to the rest of the tree are why some more recent ancestral sequences in the tree of Figure A–1 are less reconstructable than the Boreoeutherian ancestor, which acts almost like the root of a star topology.

Acknowledgements. We thank Jim Kent, Arian Smit, Adam Siepel, Gill Bejerano, Elliot Margulies, Brian Lucena, Leonid Chindelevitch, and Ron Davis for helpful discussions and suggestions. A.B.D. was supported by a NSERC post-graduate scholarship. W.M. was supported by grant HG-02238 from the National Human Genome Research Institute, E.G. was supported by NHGRI, D.H. and M.B. were supported by NHGRI Grant 1P41HG02371 and the Howard Hughes Medical Institute. Finally, we thank the NISC Comparative Sequencing Program for providing multi-species comparative sequence data.

APPENDIX B
Missing Data in phylogenetic reconstruction

This appendix presents our contribution in improving phylogenetic reconstruction accuracy obtained from dataset with partial missing data. it is published as proceeding [57].

B.1 Preface

B.2 Abstract.

The problem of phylogenetic inference from datasets including incomplete characters is among the most relevant issues in systematic biology. In this paper, we propose a new probabilistic method for estimating unknown nucleotides before computing evolutionary distances. It is developed in the framework of the Tamura-Nei evolutionary model [230]. The proposed strategy is compared, through simulations, to existing methods "Ignoring Missing Sites" (IMS) and "Proportional Distribution of Missing and Ambiguous Bases" (PDMAB) included in the PAUP package [229].

B.3 Introduction

Incomplete datasets can arise in a variety of practical situations. For example, this is often the case in molecular biology, and more precisely in phylogenetics, where an additive tree (i.e. phylogenetic tree) represents an intuitive model of species evolution. The fear of missing data often deter systematists from including in the analysis the sites with missing characters [251, 208]. Huelsenbeck [117] and Makarenkov and Lapointe [158] pointed out that the presence of taxa comprising big percentage of unknown nucleotides might considerably deteriorate the accuracy of the phylogenetic analysis. To avoid this, some authors proposed to exclude characters containing missing

data (e.g. [120] and [221]). In contrast, Wiens argued against excluding characters and showed a benefit of "filling the holes" in a data matrix as much as possible [251]. The popular PAUP software [229] includes two methods for computing evolutionary distances between species from incomplete sequence data. The first method, called IMS ("Ignoring missing sites"), is the most commonly used strategy. It proceeds by the elimination of incomplete sites while computing evolutionary distances. According to Wiens, such an approach represents a viable solution only for long sequences because of the presence of a sufficient number of known nucleotides [252]. The second method included in PAUP, called PDMAB ("Proportional distribution of missing and ambiguous bases"), computes evolutionary distances taking into account missing bases. In this paper we propose a new method, called PEMV ("Probabilistic estimation of missing values"), which estimates the identities of all missing bases prior to computing pairwise distances between taxa. To estimate a missing base, the new method proceeds by computing a similarity score between the sequence comprising the missing base and all other sequences. A probabilistic approach is used to determine the likelihood of an unknown base to be either A, C, G or T for DNA sequences. We show how this method can be applied in the framework of Tamura-Nei evolutionary model [230]. This model is considered as a further extension of the Jukes-Cantor [126], Kimura 2-parameter [132], HKY [105], and F84 [88] models. In the next section we introduce the new method for estimating missing entries in sequence data. Then, we discuss the results provided by the methods IMS, PDMAB and PEMV in a Monte Carlo simulation study carried out with DNA sequences of various lengths, containing different percentages of missing bases.

B.4 Probabilistic estimation of missing values

The new method for estimating unknown bases in nucleotide sequences, PEMV, is described here in the framework of the Tamura-Nei [230] model of sequence evolution. This model assumes that the equilibrium frequencies of nucleotides (π_A, π_C, π_G and π_T) are unequal and substitutions are not equally likely. Furthermore, it allows for three types of nucleotide substitutions: from purine (A or G) to purine, from pyrimidine (C or T) to pyrimidine and from purine to pyrimidine (respectively, from pyrimidine to purine). To compute the evolutionary distance between a pair of sequences within this model, the following formula is used:

$$\begin{aligned} D &= -\frac{2\pi_A\pi_G}{\pi_R} ln\left(1-\frac{\pi_R}{2\pi_A\pi_G}P_R-\frac{1}{2\pi_R}Q\right) \\ &\quad -\frac{2\pi_C\pi_T}{\pi_Y} ln\left(1-\frac{\pi_Y}{2\pi_C\pi_T}P_Y-\frac{1}{2\pi_Y}Q\right) \\ &\quad -\left(\pi_R\pi_Y-\frac{\pi_A\pi_G\pi_Y}{\pi_R}-\frac{\pi_C\pi_T\pi_R}{\pi_Y}\right) ln\left(1-\frac{1}{2\pi_R\pi_Y}Q\right), \end{aligned} \tag{B.1}$$

where P_R,P_Y and Q are respectively the transitional difference between purines, the transitional difference between pyrimidines and the transversional difference involving pyrimidine and purine; π_R and π_Y are respectively the frequencies of purines ($\pi_A + \pi_G$) and pyrimidines ($\pi_C + \pi_T$).

Assume that $\mathbf{C}$ is a matrix of aligned sequences, the base k,denoted as X, in the sequence i is missing and X is either A, C, G or T. To compute the distance between the sequence i and all other considered sequences, PEMV estimates, using Equation B.2 below, the probabilities $P_{ik}(X)$, to have the nucleotide X at site k of the sequence i. The probability that an unknown base at site k of the sequence i is a specific nucleotide depends on the number of sequences having this nucleotide at this site as well as on the distance (computed ignoring the missing sites) between i and all other considered sequences

having known nucleotides at site k. First, we calculate the similarity score δ between all observed sequences while ignoring missing data. For any pair of sequences, this score is equal to the number of matches between homologous nucleotides divided by the number of comparable sites.

$$P_{ik}(X) = \frac{1}{N_k}\left(\sum_{\{j|C_{jk}=X\}} \delta_{ij} + \frac{1}{3}\sum_{\{j|C_{jk}\neq X\}} (1-\delta_{ij})\right), \tag{B.2}$$

where N_k is the number of known bases at site k (i.e. column k) of the considered aligned sequences, and δ_{ij} is the similarity score between the sequences i and j computed ignoring missing sites. The following theorem characterizing the probabilities $P_{ik}(A)$, $P_{ik}(C)$, $P_{ik}(G)$ and $P_{ik}(T)$, can be stated:

Theorem 1. *For any sequence i, and any site k of the matrix C, such that C_{ik} is a missing nucleotide, the following equality holds*: $P_{ik}(A) + P_{ik}(C) + P_{ik}(G) + P_{ik}(T) = 1$.

Due to space limitation the proof of this theorem is not presented here.

Once the different probabilities P_{ik} are obtained, we can compute for any pair of sequences i and j, the evolutionary distance using Equation B.1. First, we have to calculate the nucleotide frequencies (Equation B.3), the transitional differences P_R and P_Y(Equation B.4), and the transversional difference Q (Equation B.5). Let π_X be the new frequency of the nucleotide X:

$$\pi_X = \frac{\Lambda_X^i + \sum_{\{k|C_{ik}=?\}} P_{ik}(X) + \Lambda_X^j + \sum_{\{k|C_{jk}=?\}} P_{jk}(X)}{2L}, \tag{B.3}$$

where X denotes the nucleotide A, C, G or T;Λ_X^i is the number of nucleotides X in the sequence i; symbol ? represents a missing nucleotide; L is the total

number of sites compared.

$$P(i,j) = \frac{P'(i,j) + \sum_{\{k|(C_{ik}=?orC_{jk}=?)\}} P'(i,j,k)}{L}, \tag{B.4}$$

$$Q(i,j) = \frac{Q'(i,j) + \sum_{\{k|(C_{ik}=?orC_{jk}=?)\}} Q'(i,j,k)}{L}, \tag{B.5}$$

where $P'(i,j)$ is the number of transitions of the given type (either purine to purine P'_R, or pyrimidine to pyrimidine P'_Y) between the sequences i and j computed ignoring missing sites; $P'(i,j,k)$ is the probability of transition of the given type between the sequences i and j at site k when the nucleotide at site k is missing either in i or in j (e.g. if the nucleotide at site k of the sequence i is A and the corresponding nucleotide in j is missing, the probability of transition between purines is the probability that the missing base of the sequence j is G, whereas the probability of transition between pyrimidines is 0); $Q'(i,j)$ is the number of transversions between i and j computed ignoring missing sites; $Q'(i,j,k)$ is the probability of transversion between i and j at site k when the nucleotide at site k is missing either in i or in j.

When both nucleotides at site k of i and j are missing, we use similar formulas as those described in [64]. It is worth noting that PEMV method can be used to compute the evolutionary distance independently of the evolutionary model (Equation B.6):

$$d^*_{ik} = \frac{N^c_{ij} - N^m_{ij} + \sum_{\{k|(C_{ik}=?orC_{jk}=?)\}}(1 - P^k_{ij})}{L}, \tag{B.6}$$

where N^m_{ij} is the number of matches between homologous nucleotides in the sequences i and j; N^c_{ij} is the number of comparable pairs of nucleotides in i and j (i.e. when both nucleotides are known in the homologous sites of i and j); P^k_{ij} is the probability to have a pair of identical nucleotides at site k of i and j.

B.5 Simulation study

A Monte Carlo study has been conducted to test the ability of the new method to compute accurate distances matrices that can be used as input of distance-based methods of phylogenetic analysis. We examined how the new PEMV method performed, compared to the PAUP strategies, testing them on random phylogenetic data with different percentages of missing nucleotides. The results were obtained from simulations carried out with 1000 random binary phylogenetic trees with 16 and 24 leaves. In each case, a true tree topology, denoted T, was obtained using the random tree generation procedure proposed in [139]. The branch lengths of the true tree were computed using an exponential distribution. Following the approach of Guindon and Gascuel [100], we added some noise to the branches of the true phylogeny to create a deviation from the molecular clock hypothesis. The source code of our tree generation program, written in C, is available at the following website: http://www.labunix.uqam.ca/~makarenv/tree_generation.cpp.

The random trees were then submitted to the SeqGen program [195] to simulate sequence evolution along their branches. We used SeqGen to obtain the aligned sequences of the length l (with 250, 500, 750, and 1000 bases) generated according to the HKY evolutionary model [105] which is a submodel of Tamura-Nei [230]. According to Takashi and Nei (2000), the following equilibrium nucleotide frequencies were chosen: $\pi_A = 0.15$, $\pi_C = 0.35$, $\pi_G = 0.35$, and $\pi_T = 0.15$. The transition/transversion rate was set to 4. To simulate missing data in the sequences, we used one of the two strategies described by Wiens (2003). This strategy consists of the random elimination of blocks of nucleotides of different sizes. This elimination is certainly more realistic from a biological point of view. Here, we generated data with 0 to 50% of missing bases. The obtained sequences were submitted to the three methods

for computing evolutionary distances. For each distance matrix provided by IMS, PDMAB and PEMV, we inferred a phylogeny T' using the BioNJ algorithm [93]. The phylogeny T' was then compared to the true phylogeny T

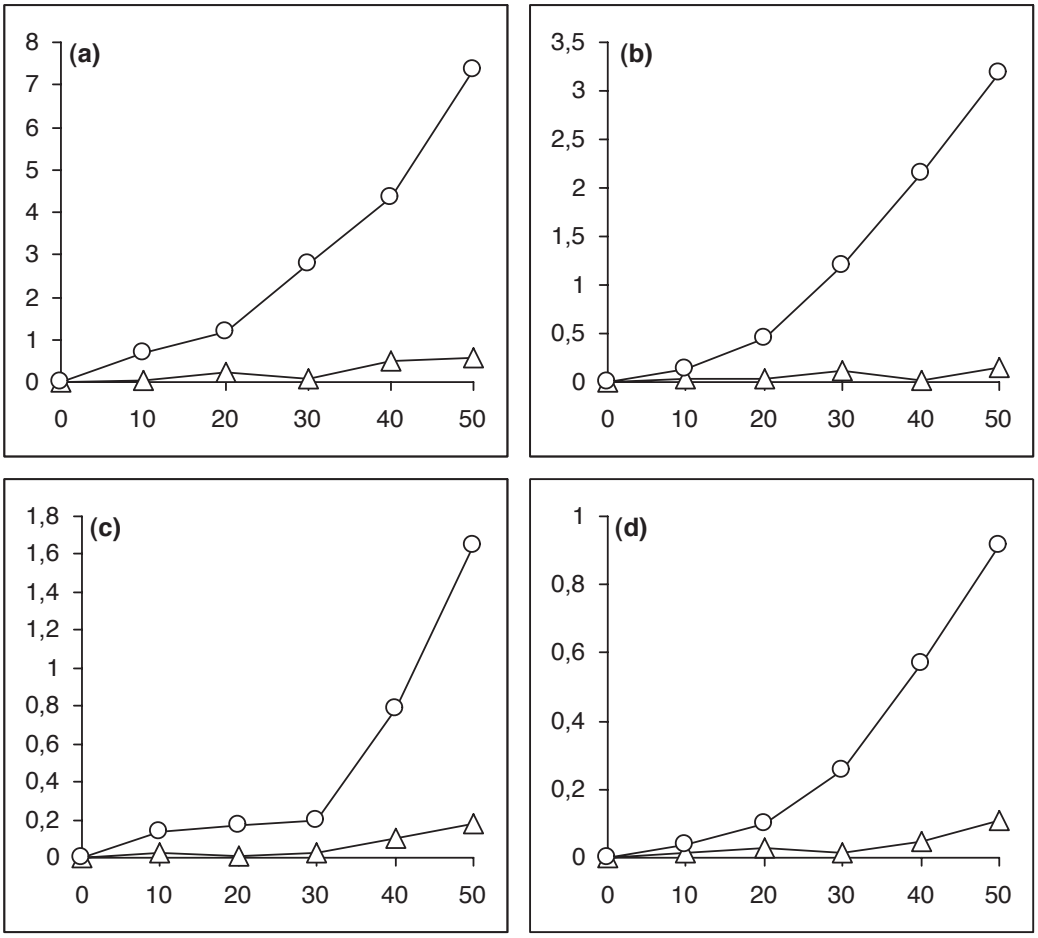


Figure B–1: Improvement in topological recovery obtained for random phylogenetic trees with 16 species. The percentage of missing bases varies from 0 to 50% (abscissa axis). The curves represent the gain (in %) against the less accurate method of PAUP. The difference was measured as the variation of the Robinson and Foulds topological distance between the less accurate method of PAUP and the most accurate method of PAUP (△) and PEMV (◯).The sequences with (a) 250 bases, (b) 500 bases, (c) 750 bases, and (d) 1000 bases are represented.

using the Robinson and Foulds topological distance [201]. The Robinson and Foulds distance between two phylogenies is the minimum number of operations, consisting of merging and splitting internal nodes, which are necessary

to transform one tree into another. This distance is reported as percentage of its maximum value ($2n$-6 for a phylogeny with n leaves). The lower this value is, the closer the obtained tree T' to the true tree T.

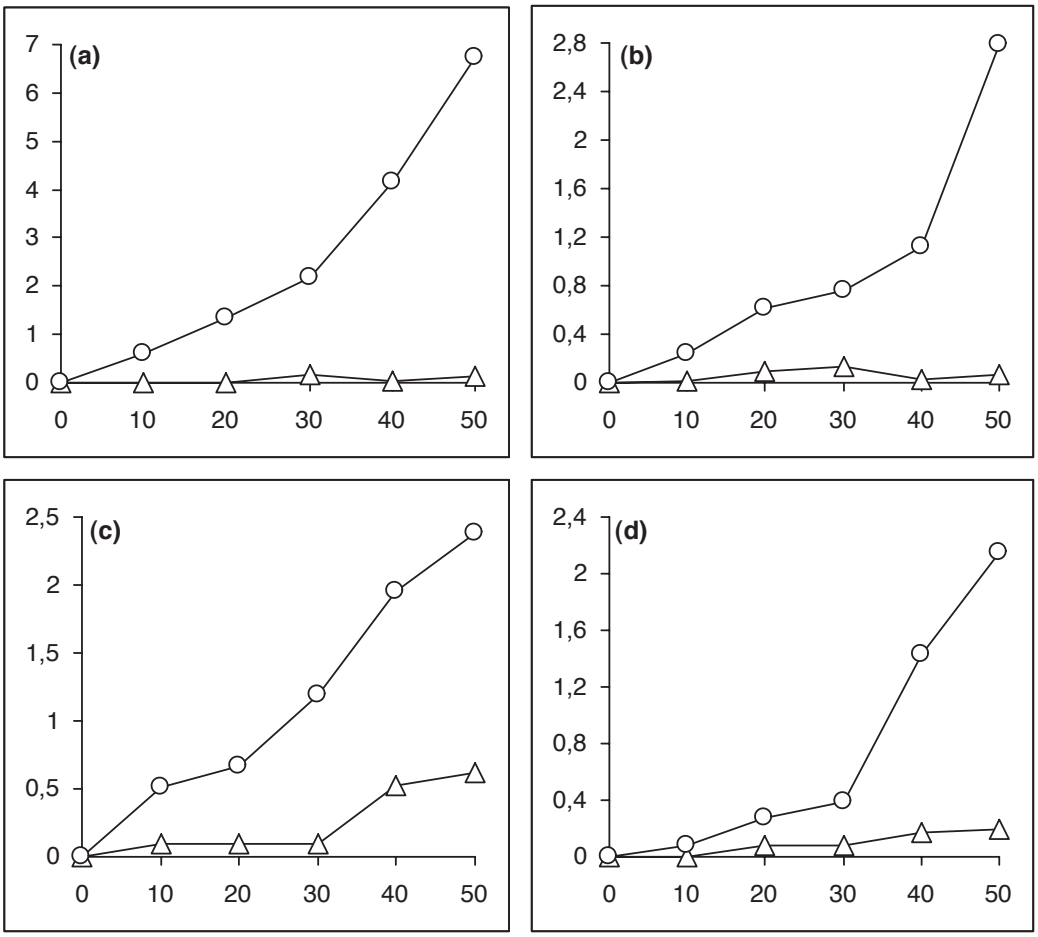


Figure B–2: Improvement in topological recovery obtained for random phylogenetic trees with 24 species. The percentage of missing bases varies from 0 to 50% (abscissa axis). The curves represent the gain (in %) against the less accurate method of PAUP. The difference was measured as the variation of the Robinson and Foulds topological distance between the less accurate method of PAUP and the most accurate method of PAUP (△) and PEMV (◯). The sequences with (a) 250 bases, (b) 500 bases, (c) 750 bases, and (d) 1000 bases are represented.

For each dataset, we tested the performance of the three methods depending on the sequence length. Figures B–1 and B–2 present the results given by the three competing methods for the phylogenies with 16 and 24 leaves. First,

for the phylogenies of both sizes PEMV clearly outperformed the PAUP methods (IMS and PDMAB) when the percentage of missing data was large (30% to 50%). Second, the results obtained with IMS were very similar to those given by PDMAB. Third, the gain obtained by our method was decreasing while the sequences length was increasing. At the same time, the following trend can be observed: the impact of missing data decreases when sequence length increases. Note that the same tendency has been pointed out by Wiens [252].

B.6 Conclusion

The PEMV technique introduced in this article is a new efficient method that can be applied to infer phylogenies from nucleotide sequences comprising missing data. The simulations conducted in this study demonstrated the usefulness of PEMV in estimating missing bases prior to phylogenetic reconstruction. Tested in the framework of the Tamura-Nei model [230], the PEMV method provided very promising results. The deletion of missing sites, as it is done in the IMS method, or their estimation using PDMAB (two methods available in PAUP) can remove important features of the data at hand. In this paper, we presented PEMV in the framework of the Tamura-Nei [230] model which can be viewed as a generalization of the popular F84 [88] and HKY85 [105] models. It would be interesting to extend and test this probabilistic approach within Maximum Likelihood and Maximum Parsimony models. It is also important to compare the results provided by BioNJ to those obtained using other distance-based methods of phylogenetic reconstruction, as for example, NJ [205], FITCH [86] or MW [159].

APPENDIX C
Algorithms for Detecting Complete and Partial Horizontal Gene Transfers: Theory and Practice

C.1 Preface

This appendix presents our contribution in the detection of horizontal gene transfer. This method was useful for the inference of the phages classification (see Chapter 6). It is published as proceeding [156].

C.2 Abstract.

We describe two methods for detecting horizontal gene transfers in the framework of the complete and partial gene transfer models. In case of a complete gene transfer model a new fast backward selection algorithm for predicting horizontal gene transfer events is presented. The latter algorithm can rely either on the metric or on the topological optimization to identify horizontal gene transfers between branches of a given species phylogeny. In case of the topological optimization, we use the well-known Robison and Foulds (RF) topological distance, whereas in case of the metric optimization, the least-squares (LS) criterion is considered. We also formulate and prove the NP-hardness of the partial gene transfer problem. Second, an efficient algorithm for predicting partial transfers, using the Gauss and Seidel optimization, is discussed. We also show how to assess the reliability of a specific gene transfer or a whole gene transfer scenario. In the application section, we apply the new algorithm to detect possible gene transfers occurred during the evolution of the gene `rpl12e`.

C.3 Introduction

Horizontal gene transfer (HGT) is a direct transfer of genetic material from one lineage to another. The understanding that horizontal gene transfer might have played a key role in biological evolution is one of the most fundamental changes in our perception of general aspects of molecular biology in recent years [66, 145, 144]. Bacteria and Archaea have sophisticated mechanisms for the acquisition of new genes through HGT which may have been favoured by natural selection as a more rapid mechanism of adaptation than the alteration of gene functions through numerous point mutations. If the donor DNA and the recipient chromosome display some homologous sequences, the donor sequences can be stably incorporated into the recipient chromosome by homologous recombination. The three main mechanisms of HGT are the following: transformation, consisting of uptake of naked DNA from the environment; conjugation, which is mediated by conjugal plasmids or conjugal transposons; and transduction, consisting of DNA transfer by phage. These transferring mechanisms can introduce sequences of DNA that display little similarity with the remaining DNA of the recipient cell [66].

There are a few ways to identify the genes that have been transferred horizontally. First, sequence analysis of the host genome may reveal areas with GC content or codon usage patterns atypical to it [141]. Second, if a sequence is found in only one organism and is absent from all other closely related organisms, it is more likely that it has been introduced horizontally into this organism rather than deleted from all the others. Third, the comparison of a morphology-based species tree or a molecular tree based on a molecule that is assumed to be refractory to horizontal gene transfer (e.g. 16S rRNA or 23S rRNA) against a phylogeny of an observed gene may reveal topological conflicts which can be explained by horizontal transfers.

Several attempts to use network-based models to depict horizontal gene transfers can be found (see for example: [245, 187, 34, 103], or [104]). A model of horizontal gene transfer that maps gene phylogenies into a species tree has been introduced by [103]. Mirkin *et al.* [167] and Hallett *et al.* [104] have developed algorithms allowing for simultaneous identification of gene duplications, gene losses, and horizontal gene transfers. The papers by Moret *et al.* [169, 178] give an overview of the network modeling in phylogenetics. In a recent paper published in the SFC2004 proceedings, [166] considered some approaches for biologically meaningful mapping of data of individual gene families into an evolutionary species tree. One approach first produces a gene tree, then maps it into the species tree, whereas the other approach first takes the gene phyletic profile, maps it into the species tree and then tunes it into a directed scenario based on the similarity data.

In this article we continue the work started in Ref. [18], where we described a HGT model based on least-squares, and in Ref. [160], where we showed the difference between complete and partial gene transfer models. First, we describe a polynomial-time HGT algorithm for the detection of complete transfers and test it with respect to the two optimization criteria: Least-squares (LS) and Robinson and Foulds (RF) topological distance. We also suggest how to assess the reliability of horizontal gene transfers identified by our algorithm. In the application section, we show how the new algorithm predicts transfers of the gene `rpl2e` for the group of 14 Archaea organisms which were originally examined in Ref. [162].

C.4 Algorithms for Predicting Horizontal Gene Transfers

C.4.1 Basic definitions

We start this section with some basic definitions about phylogenetic trees and tree metrics, generally following the terminology of Barthélemy and Guénoche [7,

8]. The distance $\delta(x, y)$ between two vertices x and y in a phylogenetic (i.e. additive) tree T is defined as the sum of the edge lengths in the unique path linking x and y in T. Such a path is denoted (x, y). A leaf is a vertex of degree one.

Definition C.1 *Let X be a finite set of n taxa. A* dissimilarity *d on X is a non-negative function on $(X \times X)$ such that for any x, y from X:*

1. $d(x, y) = d(y, x)$, *and*
2. $d(x, y) = d(y, x) \geq d(x, x) = 0$.

Definition C.2 *A dissimilarity d on X satisfies the* four-point condition *if for any x, y, z, and w from X:*

$$d(x, y) + d(z, w) \leq \text{Max}\{d(x, z) + d(y, w); d(x, w) + d(y, z)\}.$$

Definition C.3 *For a finite set X, a* ***phylogenetic tree*** *(i.e. an additive tree or a X-tree) is an ordered pair (T, ϕ) consisting of a tree T, with vertex set V, and a map $\phi : X \rightarrow V$ with the property that, for all $x \in X$ with degree at most two, $x \in \phi(X)$. A phylogenetic tree is* ***binary*** *if ϕ is a bijection from X into the leaf set of T and every interior vertex has degree three.*

The main theorem relating the four-point condition and dissimilarity representability by a phylogenetic tree (i.e., phylogeny) is as follows:

Theorem C.1 (Zarestskii, Buneman, Patrinos & Hakimi, Dobson) *Any dissimilarity satisfying the four-point condition can be represented by a phylogenetic tree such that for any x, y from X, $d(x, y)$ is equal to the length of the path linking the leaves x and y in T. This dissimilarity is called a* tree metric. *Furthermore, this tree is unique.*

Figure C–1 is an example of a tree metric on the set X of 5 taxa and the associated phylogenetic tree.

	x2	x3	x4	x5
x1	6	6	4	2
x2		2	4	6
x3			4	6
x4				4

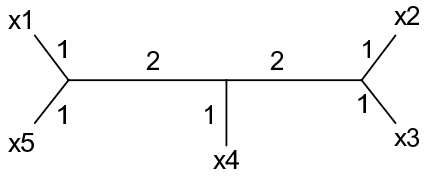


Figure C–1: An example of a tree metric on the set X of 5 taxa.

C.4.2 Optimization criteria

Here we present a fast greedy algorithm for predicting complete horizontal gene transfers. The algorithm for identifying HGTs proceeds by a progressive reconciliation of the given species and gene phylogenetic trees, denoted T and T' respectively. Usually, the species tree T is inferred from the genes that are refractory to horizontal gene transfer and genetic recombination (e.g., 16sRNA sequences). This tree represents the direct or tree-like evolution. The gene tree T' represents the evolution of a given gene which is supposed to undergo horizontal transfers.

At each step of the algorithm, all pairs of branches in T are tested against the hypothesis that a horizontal gene transfer has occurred between them. The considered HGT model assumes that the transferred gene supplants the entire homologous gene of the host or that the homologous gene is simply absent at the host genome. In such a model, the original species phylogenetic tree T is gradually transformed into the gene phylogenetic tree T' through a series of subtree moves (i.e., gene transfers or HGTs). The topology of the gene tree T' is kept fixed. The goal is to find the minimum possible sequence of trees $T, T_1, T_2, \ldots, T'$ that transforms T into T'. Obviously, a number of necessary biological rules should be taken into account. For instance, the transfers within the same lineage as well as some double-crossing transfers should be prohibited (for more detail, see [154, 188, 189, 103]).

We consider two optimization criteria which can be used at each algorithmic step to select the best HGT. The first optimization criterion that we consider is the *least-squares* (LS) *function* Q. It is computed as follows:

$$Q = \sum_i \sum_j \big(d(i,j) - \delta(i,j)\big)^2, \tag{C.1}$$

where $d(i,j)$ is the pairwise distance between the leaves i and j in the species tree T (or in the tree T_1 obtained from T after the first subtree move) and $\delta(i,j)$ the pairwise distance between i and j in the gene tree T'. The second criterion that can be useful for assessing discrepancy between the species and gene phylogenies is the Robinson and Foulds (RF) topological distance [201]. The RF metric is an important and frequently used tool to compare the topologies of phylogenetic trees. This distance is equal to the minimum number of elementary operations, consisting of merging and splitting nodes, necessary to transform one tree into the other. This distance is also the number of bipartitions or Buneman's splits belonging to exactly one of the two trees. When the RF distance is considered, we can use it as an optimization criterion as follows: all possible transformations of the species tree, consisting of transferring one of its subtrees from one branch to another, are evaluated in a way that the RF distance between the transformed species tree T_1 and the gene tree T' is computed. The subtree transfer providing the minimum of the RF distance between T_1 and T' is retained. Note that the problem asking to find the minimum number of subtree transfer operations necessary to transform one tree into another (i.e. also known as *Subtree Transfer Problem*) *has been shown to be* NP*-hard* [110].

C.4.3 Greedy backward algorithm for predicting complete horizontal gene transfers

In this section we discuss the main features of our algorithm based on the backward selection of horizontal gene transfers. Consider a gene transfer in the species tree T going from b to a and transforming it into the tree T_1 (Fig. C–2). The following timing constraint is considered (see also Ref. [160]): to allow the transfer between the branches (z, w) and (x, y) of the species tree T, the cluster combining the subtrees rooted by the vertices y and w must be present in the gene tree T'. Such a constraint enables us, first, to arrange the topological conflicts between T and T' that are due to the transfers between single species or their close ancestors and, second, to identify the transfers that have occurred deeper in the phylogeny (i.e., closer to the tree root). The usage of this constraint allows the method to follow the order that is opposite to the order of evolution and infer first the most recent HGTs which are easier to detect.

Proposition C.1 *If all bipartitions corresponding to the branches of the path* (x, z) *in the transformed species tree* T_1 (*Fig.* C–2) *can be found in the bipartition table of the gene tree* T'*, then the transfer from* b *to* a*, transforming the*

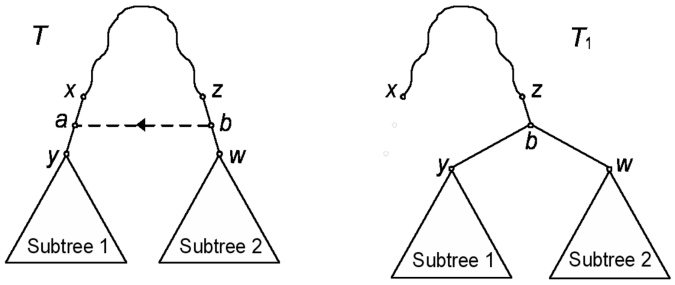


Figure C–2: Subtree constraint: the transfer between the branches (z, w) and (x, y) of the species tree T can be allowed if and only if the cluster regrouping both affected subtrees is present in the gene tree; here, a single branch is depicted by a plane line and a path is depicted by a wavy line.

species tree T into T_1, is a part of a minimum cost HGT *scenario transforming T into T'.*

This Proposition can be easily proved by induction on the number of branches of the path (x, z).

The main steps of the HGT detection algorithm are the following:

Preliminary step. Infer species and gene phylogenies, denoted respectively T and T', whose leaves are labeled by the same set of n taxa. Both species and gene trees must be rooted. If there exist identical subtrees with two or more leaves belonging to both T and T', reduce the size of the problem by replacing these subtrees with the same auxiliary taxa in both T and T'.

Step 1 ($\ldots k$) *Test all possible HGTs between pairs of branches in T_{k-1} ($T_{k-1} = T$ at Step 1) except the transfers between adjacent branches and those violating the evolutionary and subtree constraints. If no such a transfer exists, relax the subtree constraint. In our simulations described in the section Simulation study, this relaxation was necessary on average in 1.2% of cases. Search for the transfers satisfying the conditions of Proposition C.1. If no such transfers exist, choose the best HGT with respect to the selected optimization criterion that can be in our case: the least-squares (LS) or the Robinson and Foulds (RF) metric. Reduce the size of the problem by contracting the newly-formed subtree in the transformed species tree T_k and the gene tree T'. In the list of the obtained HGTs, search for and eliminate the idle transfers using a backward procedure. An idle transfer is the transfer whose removal does not change the topology of the tree T_k.*

Stopping condition and time complexity. The procedure stops when the LS or RF coefficient equals zero. Such a computation requires $O(kn^4)$ time to generate k transfers in a phylogenetic tree with n leaves. However, because

of the progressive size reduction of the species and gene trees, the practical time complexity of this algorithm is rather $O(kn^3)$.

Proposition C.2 *If the subtree constraint is not relaxed, the* HGT *detection algorithm requires at most* $n-3$ *steps to transform a binary species tree with* n *leaves into a binary gene tree with the same set of* n *leaves.*

The proof of this proposition is based on the fact that the maximum value of the RF distance between two binary trees with n leaves is $2n-6$ and that each subtree transfer satisfying the subtree constraint decreases the value of the RF distance by at least 2.

C.4.4 Partial gene transfer model

The partial gene transfer model is more general, but also more complex and challenging. It presumes that only a part of the transferred gene has been acquired by the host species through the process of homologous recombination [160]. This means that the traditional species phylogenetic tree is transformed into a directed phylogenetic network (i.e. a directed connected graph). For example, Denamur *et al.* [48] proposed a method to identify gene segments being transferred horizontally. This method was applied to detect partial HGTs of the mut$U$ and mut$S$ genes within E. coli evolutionary trees. Because many analyzes are now directed at understanding the evolution of complete genomes, the partial gene transfer model could be also useful if one wanted to model the transfer of a portion of a genome.

In a phylogenetic tree, there is always a unique path connecting a pair of nodes. Adding to it a HGT branch creates an extra path between certain nodes. Figure C–3 illustrates the case where the evolutionary distance between the taxa i and j can be affected by the addition of the HGT branch (b, a) representing partial gene transfer from b to a. It is relevant to assume that the HGT from b to a can affect the evolutionary distance between the taxa

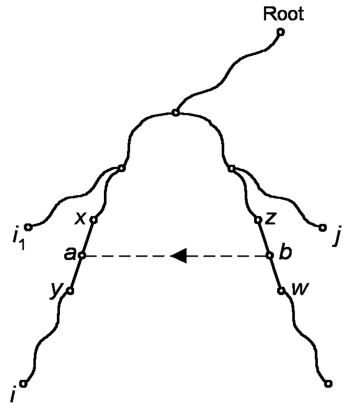


Figure C–3: Evolutionary distance between the taxa i and j can be allowed to change after the addition of the branch (b, a) representing a partial HGT between the branches (z, w) and (x, y). Evolutionary distance between the taxa i_1 and j must not be affected by the addition of (b, a).

i and j if and only if the destination point a is located on the path between i and the root of the tree; the position of j is fixed. Thus, in the reticulate phylogeny T in Fig. C–3 the evolutionary distance $d_1(i, j)$ between the taxa i and j can be computed as follows:

$$d_1(i, j) = (1 - \alpha)d(i, j) + \alpha\big(d(i, a) + d(j, b)\big), \tag{C.2}$$

where α indicates the fraction, unknown in advance, of the transferred gene and d is the internode distance in the species tree before the addition of the HGT branch (b, a).

On the contrary, the distance between the taxa i_1 and j (Fig. C–3) must not be affected by the addition of (b, a). Figure C–4 illustrates the other cases where the addition of a HGT branch must not affect the length of the evolutionary path between i and j.

The least-squares loss function Q to be minimized with the *unknown vector of edge lengths* ℓ in T and the *unknown fraction of the transferred gene*

α is as follows:

$$Q(L,\alpha) = \sum_{ij\in S}\left((1-\dot{a})\sum_{k\in \mathrm{path}(ij)}\ell_{ij}^{k} + \alpha\left(\sum_{k\in \mathrm{path}(ia)}\ell_{ia}^{k} + \sum_{k\in \mathrm{path}(jb)}\ell_{jb}^{k}\right) - \delta(i,j)\right)^{2} + \sum_{ij\notin S}\left(\sum_{k\in \mathrm{path}(ij)}\ell_{ij}^{k} - \delta(i,j)\right)^{2} \longrightarrow \min, \quad \text{(C.3)}$$

where $\delta(i,j)$ is the given gene dissimilarity between i and j; ℓ_{ij}^{k} is the length of the branch k of the path (ij) in T; α is the fraction of the transferred gene $(0 \leq \alpha \leq 1)$; and S is the set of pairs of taxa $\{ij\}$ such that the transfer (ba) can affect the evolutionary distance between them.

To show the NP-hardness of the least-squares optimization in the context of the partial gene transfer the following problem can be stated:

Given: Species phylogenetic tree T (with the associated tree metric d on the set of taxa X), gene dissimilarity $\boldsymbol{\delta}$ on X, and a fixed non-negative value ε.

Find the minimum number of partial gene transfers k such that:

$$Q = \sum_{i}\sum_{j}\left(d_k(i,j) - \delta(i,j)\right)^2 \leq \varepsilon, \quad \text{(C.4)}$$

where $d_k(i,j)$ is the network distance between i and j, computed using Formulae C.2 and C.3, in the phylogenetic network T_k obtained from T after the addition of k partial gene transfers.

Theorem C.2 *The minimum number of partial gene transfer problem (MNPGT problem) is* NP*-hard.*

The proof of this theorem is based on a polynomial-time reduction from the *Subtree Transfer Problem* (STR problem) that consists of finding the minimum number of complete gene transfers to transform a given species tree

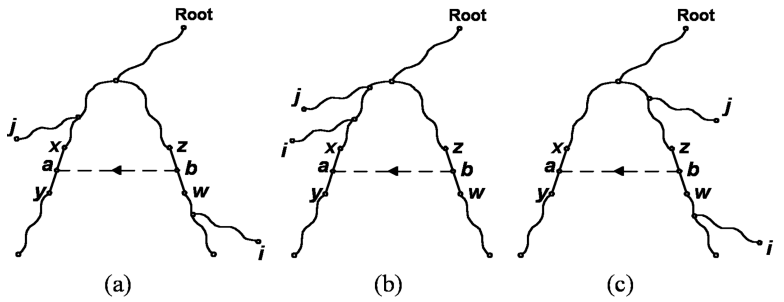


Figure C–4: Three situations when the evolutionary distance between the taxa i and j must not be affected by the addition of the new branch (b, a) representing a partial HGT between the branches (z, w) and (x, y). Path between the taxa i and j cannot to go through the branch (b, a).

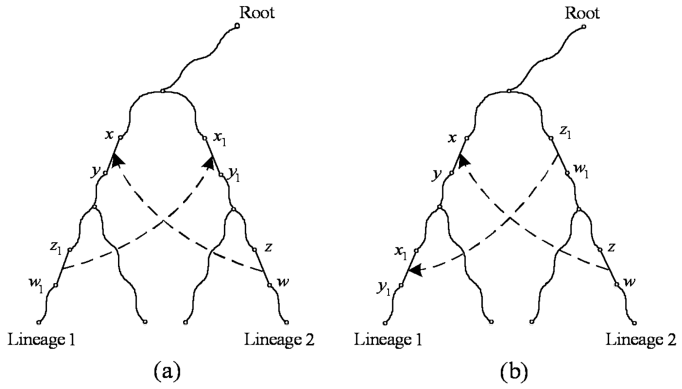


Figure C–5: Transfers between two lineages crossing in such ways must be prohibited.

T into a given gene tree T'. The STR problem is identical to the problem of adding to T the minimum number of complete gene transfers such that $Q = \sum_i \sum_j \left(d_k(i,j) - \delta(i,j)\right)^2 \leq 0$ (i.e., the case of $\varepsilon = 0$ is considered), where $d_k(i,j)$ is the pairwise distance between i and j in the phylogenetic tree (i.e., a particular case of a phylogenetic network). Here, the tree T_k is obtained from T after the addition of k complete gene transfers (i.e., a particular case of a partial transfer) and $\delta(i,j)$ is the given tree metric associated with T'.

Several important timing constraints have to be incorporated into this model, in addition to those taken into account in the complete HTS model, to identify the interactions between HGTs that are not intelligible from an

evolutionary point of view. Some of these constraints, but not all of them, were initially pointed out by Page and Charleston [188, 189]. For instance, double-crossing transfers between two lineages (Figs. 5a and b) must be forbidden. In this case, the HGT events affect the ancestor of the species from the previous transfer. Making the source and destination lineages contemporaneous for one HGT makes the other transfer impossible (Fig. C–5).

Note that the rule illustrated in Figure 5a is automatically taken into account in the complete gene transfer model, where its violation would be equivalent to the violation of the same lineage constraint (see Page and Charleston [188, 189]). For instance (Figure 5a), the HGT from (z, w) to (x, y) cannot be followed by the transfer from (z_1, w_1) to (x_1, y_1) because after the first HGT the branches (z_1, w_1) and (x_1, y_1) will be located on the same lineage (Lineage 2). We also identify two cases, where the evolutionary distance between the taxa i and j can be affected by multiple transfers (Figures 6a and b); and, two cases, where this distance must not be affected by them (Figures 6c and d). Failure to take these constraints into account can result in postulating transfers that are mutually incompatible.

Assume that a partial gene transfer between the branches (z, w) and (x, y) (i.e., from b to a in Fig. C–3) of the species tree T has taken place. The lengths of all branches in T are reassessed in the least-squares sense after the addition of (b, a), whereas the length of (b, a) is assumed to be 0. To reassess the branch lengths of T, we have first to make an assumption about the value of the parameter α (eq. C.2), indicating the gene fraction being transferred. This parameter can be estimated either by comparing sequence data corresponding to the subtrees rooted by the vertices y and w, or different values of α can be tested in the optimization problem.

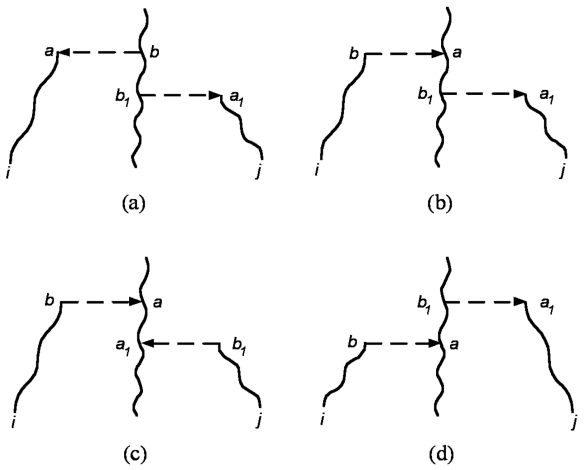


Figure C–6: Cases (a) and (b): evolutionary path between the taxa i and j can go through both HGT branches (b, a) and (b_1, a_1). Cases (c) and (d): evolutionary path between the taxa i and j cannot go through both HGT branches (b,a) and (b_1, a_1).

Fixing the parameter α, we reduce to a linear system the system of equations establishing the correspondence between the experimental gene distances and the path-length distances in the HGT network. This system having generally more variables (i.e. branch lengths of T) than equations (i.e. pairwise distances in T; the number of equations is always $n(n-1)/2$ for n taxa) can be solved by approximation in the least-squares sense. Let us now show how the approximation problem can be stated and efficiently solved.

Let $\mathbf{A}_\alpha$ be the matrix of dimension $n(n-1)/2 \times m$, each row of which is associated with one pair of taxa of X, where n is the number of taxa and m is a number of edges in T. The value $a_{ij,e}$ of this matrix corresponding to the pair of taxa ij and the edge e is equal either to 1, or to α, or to $1-\alpha$ if the edge e is in the path (ij) in T, and is equal to 0 if not. Let ℓ be the vector of edge lengths of m elements and $\mathbf{d}$ be given vector of gene distances of $n(n-1)/2$ elements.

Fixing the value of α (e.g., values 0, 0.1, 0.2, ..., and 1.0 can be tested in turn), we obtain a linear system of $n(n-1)/2$ equations with m unknowns: $\mathbf{A}_\alpha \times \ell = \mathbf{d}$.

When $n \geq 4$, this system has more equations than unknowns. It can be solved by approximation in the least-squares sense:

$$(\mathbf{A} \times \ell - \mathbf{d})^2 \rightarrow \min . \tag{C.5}$$

After taking the gradient we have:

$$\mathbf{A}_\alpha^t \times (\mathbf{A}_\alpha \times l - \mathbf{d}) = 0. \tag{C.6}$$

Following algebraic manipulations, we obtain:

$$\mathbf{A}_\alpha^t \times \mathbf{A}_\alpha \times l = \mathbf{A}_\alpha^t \times \mathbf{d}. \tag{C.7}$$

Thus, we have: $\mathbf{B} \times \ell = \mathbf{c}$, where $\mathbf{B}$ is a $(m \times m)$ matrix, and $\mathbf{c}$ is a vector with m components.

Following Barthélemy and Guénoche [7] and Makarenkov and Leclerc [159], we apply a slightly modified Gauss-Seidel method to solve the above system. The method consists of decomposing $\mathbf{B}$ into its diagonal ($\mathbf{\Delta}$), its strictly upper triangular component ($-\mathbf{F}$), and its strictly lower triangular component ($-\mathbf{E}$):

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \dots & \dots & \dots & \dots \\ b_{m1} & b_{m2} & \dots & b_{mm} \end{pmatrix} = \begin{pmatrix} & & & -\mathbf{F} \\ & \mathbf{\Delta} & & \\ -\mathbf{E} & & & \end{pmatrix} = \mathbf{\Delta} - \mathbf{E} - \mathbf{F}. \tag{C.8}$$

Then, we apply the iterative procedure:

$$\mathbf{\Delta} \times \ell^{k+1} = \mathbf{E} \times \ell^{k+1} + \mathbf{F} \times \ell^{(k)} + \mathbf{c}, \tag{C.9}$$

which allows us to compute gradually the components of the vector $\ell(j)^{(k+1)}$, corresponding to the edge lengths at the $k+1$th iteration, from those of $\ell(j)^k$. If the computed value of $\ell(j)^{(k+1)}$ is negative, it is replaced with the value 0. This operation is equivalent to the projection on the cone $\mathbf{L} \geq 0$, which ensures an appropriate solution.

The exact equation used in this method is the following for all $j = 1, 2, \ldots, m$:

$$\ell(j)^{(k+1)} = \left(- \left(\sum_{j+1 \leq i \leq m} b_{ij} \ell(j)^{(k)} \right) - \left(\sum_{1 \leq i \leq j-1} b_{ij} \ell(j)^{(k+1)} \right) + c_j \right) \Big/ b_{jj}. \quad \text{(C.10)}$$

Thus, the main steps of the partial gene transfer algorithm can be stated as follows:

Preliminary step. This step corresponds to the preliminary step discussed in the context of the complete gene transfer model. It consists of inferring the species and gene phylogenies denoted respectively T and T' whose leaves are labeled by the same set X of n taxa. Because the classical Robinson and Foulds distance is defined only for tree topologies, we use the least-squares as a unique optimization criterion when modeling partial HGTs.

Step 2 *Test all connections between pairs of branches in the species tree* $\mathbf{T}$. *For each HGT connexion satisfying evolutionary constraints, carry out the following optimization:*

[(a)]Fix the value of the fraction of the gene being transferred α *(e.g., one can try in turn the values of 0, 0.1, 0.2, . . . , and 1.0). Compute using the Gauss–Seidel method the optimal lengths* $\mathbf{l}$ *of the edges in the species tree (or network, starting from Step 2)* $\mathbf{T}$. *Go back to the original equation system:* $\mathbf{A}_\alpha \times \mathbf{l} = \mathbf{d}$. *Fix the values of the vector* $\mathbf{l}$ *found using the Gauss-Seidel method and solve this problem by least-squares considering*

as unknown the parameter α. Then, fix the optimal value of α found and repeat the computation until both unknown parameters l and α converge to a certain solution.

All eligible pairs of branches in T can be processed in this way. The HGT connection providing the smallest value of the LS coefficient Q and satisfying the defined evolutionary constraints should be selected for the addition to the species tree T, transforming it into a phylogenetic network.

2. **Step 3** $(2, \ldots, k)$ *Run the algorithm until a fixed number k of partial gene transfers is found and added to T or the value of the LS criterion Q is lower than a pre-established threshold ε.*

Time complexity of this algorithm is $O(kn^5)$ to add k partial horizontal gene transfers to the species tree with n leaves.

C.4.5 Bootstrap validation of horizontal gene transfers

Bootstrap analysis can be used to place confidence intervals on internal branches of evolutionary trees [84]. We designed a bootstrap validation procedure for computing the bootstrap scores either for a specific gene transfer or a whole gene transfer scenario. The following strategy was adopted to assess the reliability of obtained HGTs. Because we are mostly interested in the evolution of a given gene or a group of genes, the sequences used to build the species tree are not resampled. The species tree is taken as an *a priori* assumption of the method and held constant. The sequence data used to build the gene tree are drawn with replacement in order to create a series of pseudo-replicates. The HGT detection algorithm is then carried out on the bootstrapped pseudo-replicates. Thus, for all HGT branches appearing in the original scenario, we verify if they appear in the obtained transfer scenarios, using as input the original species tree and the gene tree inferred from the sets of pseudo-replicates. It is worth noting that among resampled datasets

only those that give rise to a gene phylogenetic tree such that it contains the root branch separating this tree into exactly the same bipartition sets as the root branch of the original gene tree does, are eligible for the HGT bootstrap analysis.

Simulation study. A Monte Carlo study was conducted to test the ability of the new method to recover correct gene transfers. In the framework of *the complete* HGT *model only* we examined how the detection procedure performed depending on the model of sequence evolution, number of observed species, and sequence length. The results illustrated in Figs. C–7 and C–8, and reported in Tables C–1 and C–2 (see Appendix) were obtained from simulations carried out with random binary phylogenetic trees with 8, 16, 24, 32, 48, and 64 leaves, whereas the sequence length varied from 125 to 1000 sites. The simulation procedure consisted of the five basic steps described below:

1. A true tree topology, denoted T, was obtained using the random tree generation procedure proposed by Kuhner and Felsenstein [139]. The branch lengths of T were computed using an exponential distribution. Following the approach of Guindon and Gascuel [100], we added some noise to the branches of the true phylogenies to create a deviation from the molecular clock hypothesis. All the branch lengths of T were multiplied by $1+\alpha x$, where the variable x was obtained from a standard exponential distribution $\big(P(x > k) = \exp(-k)\big)$, where the constant a was a tuning factor for the deviation intensity. Following Guindon and Gascuel [100], a was fixed to 0.8. The random trees generated by this procedure are chosen to have the depth of $O\big(\log(n)\big)$, where n is the number of species (i.e. number of leaves in a binary phylogenetic tree).

2. Each random phylogeny was then submitted to the SeqGen program [195] to simulate sequence evolution along its branches according to the Jukes and Cantor [126], Kimura 2-parameter [132], and Jin–Nei Gamma[125] models.

3. To assess the quality of HGT detection by the new method, we developed a simulation program using the results of SeqGen. For each considered rooted tree, viewed as an organismal phylogeny, our program created one random horizontal gene transfer that respected the evolutionary constraints discussed in the algorithmic section. During this operation, the program regenerated the DNA sequences for each tree node located in the subtree affected by the HGT. As the simulations were carried out for the complete gene transfer model, the HGT destination sequence was set identical to the source sequence and the new sequences were regenerated from it according to the selected evolutionary model.

4. The sequence to distance transformation corresponding to the considered model of evolution was then applied to the DNA sequences associated with the leaves of the phylogeny affected by the gene transfer. The NJ method [205] was used to infer the gene trees from the obtained distance matrix. The topology of the organismal phylogeny (i.e. true tree T) was supposed to be known.

5. The HGT detection method was then carried out to infer the transfer. The experiments were conducted using the procedures based on the RF and LS optimization. The simulations were carried out for 500 random rooted phylogenies with 8 and 16 leaves and 100 random rooted phylogenies with 24 to 64 leaves.

Figures C–7 and C–8 present the average simulation results obtained for random phylogenies with 8 to 64 leaves, using as optimization criteria the RF topological distance and LS function, respectively. These figures illustrate

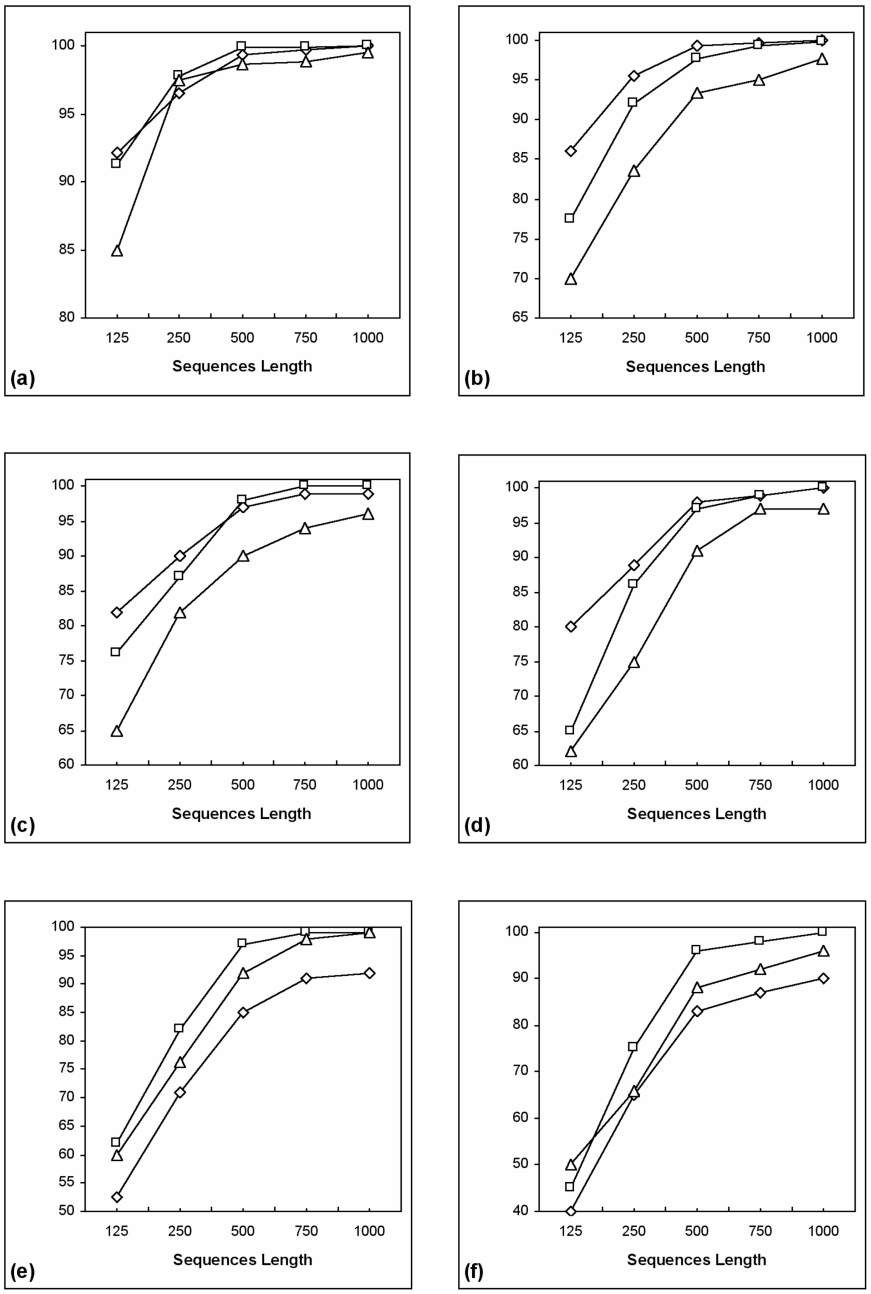


Figure C–7: HGT detection rates obtained for random phylogenies with 8 to 64 leaves ($8-a$, $16-b$, $24-b$, $32-d$, $48-e$, $64-f$) using the RF topological distance for optimization. Jukes and Cantor ($\Diamond$), Kimura 2-parameter ($\Box$), and Jin–Nei Gamma (Δ) models were used for the tree generation.

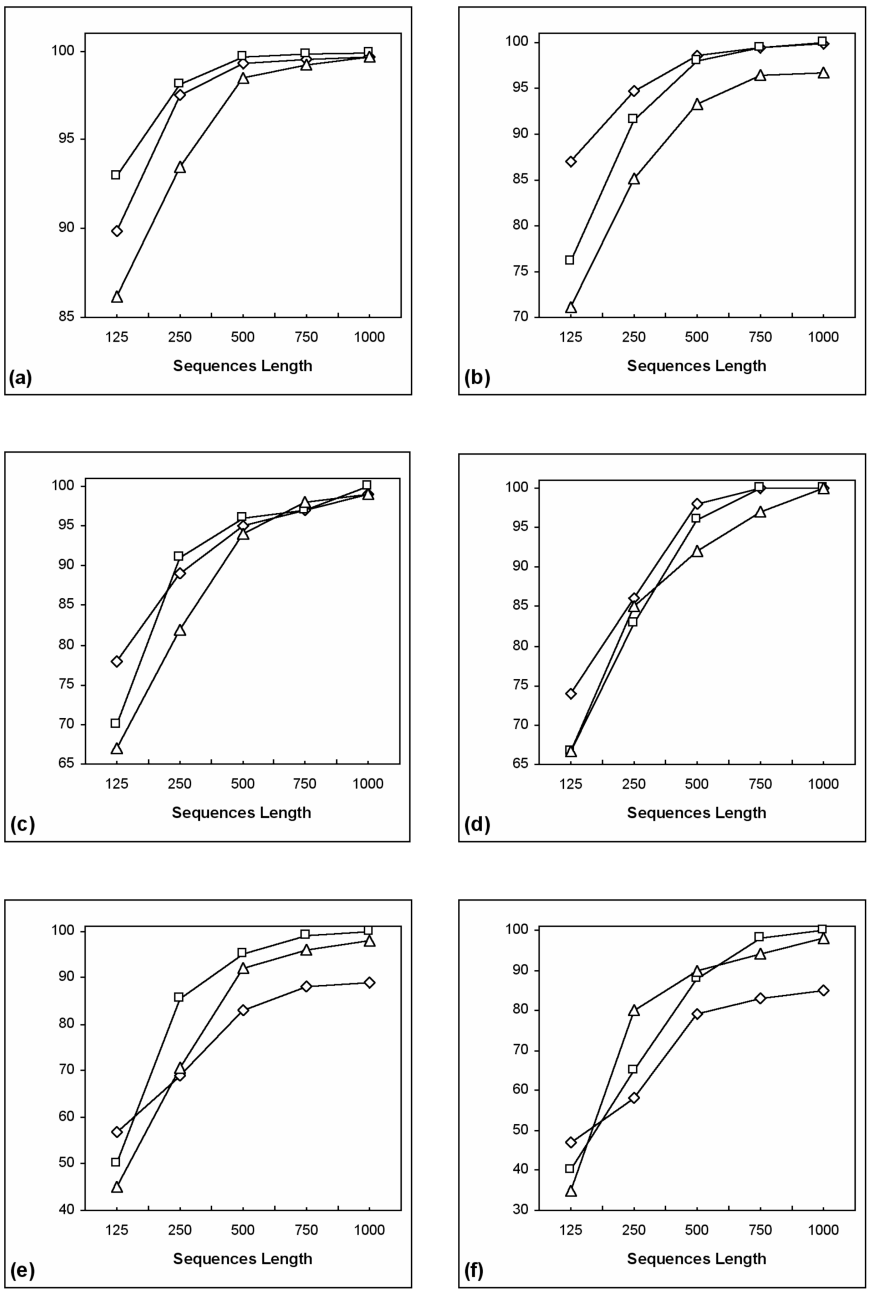


Figure C–8: HGT detection rates obtained for random phylogenies with 8 to 64 leaves ($8-a$, $16-b$, $24-b$, $32-d$, $48-e$, $64-f$) using the LS function for optimization. Jukes and Cantor ($\Diamond$), Kimura 2-parameter ($\Box$), and Jin–Nei Gamma (Δ) models were used for the tree generation.

how the detection rate changes as the number of sites varies from 125 to 1000. As expected, the detection rate grows as the number of sites increases and the number of species decreases. Note that for the phylogenies with 8 to 32 leaves the best results were obtained under the Kumura and Jukes–Cantor models. For the phylogenies with 48 to 64 species the best performances were regularly obtained under the Kimura model, whereas the results found under the Jukes–Cantor model were the worst of the three evolutionary models.

This trend can be observed in the case of both optimization criteria. Obviously, with the short sequences we have a bigger phylogenetic error that can either appear like a HGT, when it does not occur, or disguise a real HGT. Tables C–1 and C–2 (see Appendix) report the false positive and false negative (indicated in parentheses) detection rates obtained using as optimization criteria the RF distance and LS function, respectively. A false positive HGT is an incorrect transfer found by the algorithm and a false negative HGT is the right transfer that has not been detected. A false positive HGT will always occur if the gene tree inferred by NJ (see Step 4 above) is different from the true gene tree (see Step 3 above), but it can also take place when both trees are identical but a transfer going to the direction opposite to the correct HGT disguises it, leading to the same gene tree (see [154]).

False negative HGTs are mostly due to the error of inferring the gene tree, but can also happen when a transfer going to the opposite direction disguises the correct HGT. As defined, the false positive detection rate is always bigger or equal to the negative one. The analysis of Tables C–1 and C–2 shows that the false negative rate is almost as big as the false positive rate when the tests were conducted with large phylogenies (48 and 64 species) and short sequences (125 and 250 sites). The false negative rate was noticeably lower than the false

positive one in the case of the large phylogenies and long sequences. Furthermore, we have measured the recovery rates for the HGT source, destination, and source and destination combined (i.e. the latter parameter corresponds to the detection rate depicted in Figs. C–7 and C–8). These tests were carried out under the Jukes and Cantor model of sequence evolution and using the RF distance for the algorithmic optimization. Note that the transfer destinations were generally better detectable than their sources. The difference in the source-destination detection was more important for the short sequence. For example, for the sequences with 125 sites it varied, on average, from 6% (for 8 species) to 1% (for 64 species). However, for the longer sequences the source and destination rates were very similar.

Generally, the procedure based on the RF distance provided better results than that based on the LS function. Nevertheless, some noticeable exceptions (e.g. under the Kimura model for the phylogenies with 8 leaves or under the Jin–Nei model in the case of the short sequences) can be pointed out. The simulation study suggested that the accuracy of the transfer detection is highly dependable on the model of sequence evolution, number of considered species, and length of observed sequences.

C.5 Results and discussion

We first tested our algorithm on the phylogeny of 14 species of Archaea originally considered by Matte-Tailliez *et al* [162]. The latter authors discuss problems encountered when reconstructing some parts of the archaeal phylogeny, pointing out the evidence of HGT events perturbing the evolution of a number of considered genes. Matte-Tailliez *et al.* inferred the maximum likelihood tree (Figure C–10, undirected lines) based on the concatenated 53

ribosomal proteins (7,175 positions) and compared it to the maximum likelihood phylogeny of the gene `rpl2e` (Figure C–9) built for the same 14 organisms. The calculations of the best ML tree and its branch lengths for the 53 concatenated proteins were conducted using the PUZZLE program with Γ-law correction.

Given the topological incongruence of the obtained phylogenies, the authors hypothesized a few cases of lateral transfers of the gene `rpl2e`. More precisely, the case of the transfer between the clades of *Thermoplasmatales* (*Ferroplasma acidarmanus* and *Thermoplasma acidophilum*) and *Crenarchaeota* (*Aeropyrum pernix*, *Pyrobaculum aerophilum* and *Sulfolobus solfataricus*) was indicated as the most evident one.

In order to apply our method, we first reconstructed from the original sequences the topologies of the gene (Figure C–9) and species trees (Figure C–10, undirected lines). The computations were conducted in the framework of the complete gene transfer model, using the RF optimization and subtree constraint options (Figure C–2). Five directed branches needed to reconcile the species and gene topologies have been found (Figure C–10). The connection representing the transfer between the cluster of *Halobacterium sp.* and *Haloarcula marismortui* and the species *Methanobacterium thermoautotrophicum* was found in the first iteration. This transfer provided the biggest drop of the RF distance between the species and gene phylogenies; its bootstrap score is 55%.

In the second and third iterations, we found the reconciliation branches between the species *Pyrococcus horikoshii* and *Pyrococcus furiosus* and between *Sulfolobus solfataricus* and *Pyrobaculum aerophilum*. Both of these reconciliation branches link closely related species. Such kind of connections may be due to HGT as well as to local topological rearrangements necessary because

of the tree reconstruction artifacts (e.g. attraction of long branches, unequal evolutionary rates, etc). The transfer branches 4 and 5 linking the cluster of *Crenarchaeota* to the species *Thermoplasma acidophilum* and *Ferroplasma acidarmanus* can be interpreted as HGT events that might have taken place between *Thermoplasmatales* and *Crenarchaeota*.

In the second and third iterations, we found the reconciliation branches between the species *Pyrococcus horikoshii* and *Pyrococcus furiosus* and between *Sulfolobus solfataricus* and *Pyrobaculum aerophilum*. Both of these reconciliation branches link closely related species. Such kind of connections may be due to HGT as well as to local topological rearrangements necessary because of the tree reconstruction artifacts (e.g. attraction of long branches, unequal evolutionary rates, etc). The transfer branches 4 and 5 linking the cluster of *Crenarchaeota* to the species *Thermoplasma acidophilum* and *Ferroplasma acidarmanus* can be interpreted as HGT events that might have taken place between *Thermoplasmatales* and *Crenarchaeota*.

Note, that HGT between these two groups was also predicted by Matte-Taillez *et al* [162]. In fact, the transfers 4 and 5 could consist of a unique transfer between the clades of *Thermoplasmatales* and *Crenarchaeota* that was separated into two transfers by our method due to the application of the subtree constraint (Figure C–2) and the presence of the tree reconstruction artifacts. Figure C–11 illustrates the evolution of the newly formed *Thermoplasmatales-Crenarchaeota* clade involving the HGTs 4 and 5. The usage of the LS criterion instead of RF leads to the solution consisting of 6 HGTs including all transfers from Figure C–10 except the HGT number 2 that goes in the opposite direction. Note that a new reconciliation branch found with LS brings the species *Methanococcus jannaschii* to the cluster of 4 species including *Archaeoglobus*

fulgidus. This reconciliation branch turns out to be useless and have a low bootstrap score of 14%.

C.6 Conlusion

We presented two polynomial-time algorithms for detecting horizontal gene transfer events. We considered the complete and partial gene transfer models, implying at each step, either the transformation of a species phylogeny into another tree or its transformation into a network structure. The algorithm for inferring complete gene transfers exploits the discrepancies between the species and gene phylogenies either to map the gene tree into the species tree by least-squares or to compute a topological distance between them and then estimate the possibility of a HGT event between each pair of branches of the species phylogeny. The models based on the optimization of the least-squares function and the Robinson and Foulds topological distance were introduced.

Inferred HGTs should be carefully analyzed using all available information about the data in hand in order to select the transfers that will be represented as a final solution. Each gene transfer branch added to the species phylogeny aids to resolve a conflict between it and the gene tree (i.e. helps to reconcile the species and gene phylogenies). A bootstrap validation procedure allowing one to assess the reliability of a specific gene transfer or whole gene transfer scenario was proposed. A comprehensive Monte Carlo study was carried out to test the ability of the new method to recover correct HGTs. It provided very encouraging results especially when the Robinson and Foulds distance was used as an optimization criterion. The example of the evolution of the gene `rpl2e` was considered in the application section. More simulation work is required to investigate the properties of the algorithm intended to infer partial gene transfers.

As any method of phylogenetic inferring, the new HGT detection method is subject to a number of artifacts which generally affect phylogenetic analysis; the main of them being: attraction of long branches, unequal evolutionary rates, and situations when the occurrence of some HGT events almost coincides with speciation events located closely to the recipient species. It is important to investigate in greater details the impact of these artifacts on the HGT detection technique introduced in this article. It would be also interesting to extend the presented model to the case, where the gene and species trees have different numbers of taxa; this situation can take place when some species have more than one copy of the gene under consideration.

The software implementing the new algorithms for detecting complete and partial horizontal gene transfers is freely available at the following URL address: *<http://www.info2.uqam.ca/~boca05/software/>* (this is a consol version running on the Unix and Windows platforms; it is distributed along with its `C++` source code). A graphical version of this program has been also implemented and included in the *T*-Rex web server [155] at the following URL: *<http://www.trex.uqam.ca>*.

C.7 Appendix

This Appendix includes the results of the tests described in the section Simulation Study. The results reported in Tables C–1 and C–2 correspond to the graphics represented in Figures C–7 (optimization using the RF distance) and C–8 (optimization using the LS function). They were obtained from simulations carried out for random binary phylogenies with 8, 16, 24, 32, 48, and 64 leaves, whereas the sequence length varied from 125 to 1000 sites. Note that the sum of the HGT detection rate shown in Figures C–7 and C–8 and of the false negative detection rate reported in Tables C–1 and C–2 is always 100%.

Table C–1: False positive and false negative (in parentheses) detection rates obtained for random phylogenies with 8 to 64 leaves using the RF distance as an optimization criterion. A false positive HGT is an incorrect transfer found by the algorithm and a false negative HGT is the right transfer that has not been found. For each sequence length, the simulations were carried out for 500 random phylogenies with 8 and 16 leaves and 100 random phylogenies with 24 to 64 leaves.

RF rates (in %)			Sequence length				
			125	250	500	750	1000
Species number	8	Jukes-Cantor	14.9(7.8)	5.9(3.5)	1.1(0.7)	0.3(0.3)	0.0(0.0)
		Kimura	12.9(8.7)	3.3(2.2)	0.2(0.1)	0.1(0.1)	0.0(0.0)
		Jin-Nei	20.1(15.0)	3.9(2.5)	1.6(1.3)	1.1(1.1)	0.5(0.5)
	16	Jukes-Cantor	25.7(14.0)	7.1(4.5)	1.2(0.7)	0.4(0.3)	0.0(0.0)
		Kimura	35.1(22.5)	11.9(7.9)	3.2(2.3)	0.6(0.6)	0.1(0.0)
		Jin-Nei	43.0(30.0)	22.5(16.5)	7.6(6.6)	5.3(4.9)	2.3(2.3)
	24	Jukes-Cantor	36(18)	15(10)	4(3)	1(1)	1(1)
		Kimura	43(24)	24(13)	4(2)	2(0)	0(0)
		Jin-Nei	55(35)	33(18)	19(10)	9(6)	5(4)
	32	Jukes-Cantor	37(20)	29(11)	4(2)	1(1)	1(0)
		Kimura	60(35)	31(14)	8(3)	3(1)	2(0)
		Jin-Nei	70(38)	47(25)	16(9)	8(3)	8(3)
	48	Jukes-Cantor	65(48)	49(29)	28(15)	1(1)	1(0)
		Kimura	55(38)	46(18)	9(3)	3(1)	2(0)
		Jin-Nei	70(40)	58(24)	19(8)	8(3)	8(3)
	64	Jukes-Cantor	70(60)	45(35)	27(17)	23(13)	20(10)
		Kimura	65(55)	35(25)	14(4)	12(2)	10(0)
		Jin-Nei	60(50)	44(34)	22(12)	18(8)	14(4)

Table C–2: False positive and false negative (in parentheses) detection rates obtained for random phylogenies with 8 to 64 leaves using the LS function as an optimization criterion. A false positive HGT is an incorrect transfer found by the algorithm and a false negative HGT is the right transfer that has not been found. For each sequence length, the simulations were carried out for 500 random phylogenies with 8 and 16 leaves and 100 random phylogenies with 24 to 64 leaves.

RF rates (in %)			Sequence length				
			125	250	500	750	1000
Species number	8	Jukes-Cantor	17.2(10.1)	5.0(2.5)	0.8(0.7)	0.8(0.5)	0.3(0.3)
		Kimura	10.8(7.0)	2.8(1.9)	0.3(0.3)	0.2(0.2)	0.1(0.1)
		Jin-Nei	18.6(13.8)	7.8(6.5)	1.7(1.5)	0.9(0.8)	0.5(0.3)
	16	Jukes-Cantor	25.5(13.0)	7.6(5.3)	2.2(1.4)	0.8(0.5)	0.1(0.1)
		Kimura	37.6(23.8)	11.9(8.4)	2.3(2.0)	0.6(0.6)	0.0(0.0)
		Jin-Nei	40.9(28.8)	20.9(14.8)	8.1(6.7)	3.8(3.6)	3.3(3.3)
	24	Jukes-Cantor	43(22)	13(11)	5(5)	3(3)	1(1)
		Kimura	59(30)	26(9)	7(4)	4(3)	1(0)
		Jin-Nei	67(33)	26(18)	12(6)	6(2)	3(1)
	32	Jukes-Cantor	47(26)	21(14)	5(2)	0(0)	0(0)
		Kimura	56(33)	31(17)	9(4)	0(0)	0(0)
		Jin-Nei	50(33)	31(15)	12(8)	11(3)	4(0)
	48	Jukes-Cantor	53(43)	38(31)	33(7)	22(12)	19(11)
		Kimura	60(50)	34(14)	16(5)	5(1)	2(0)
		Jin-Nei	65(55)	50(29)	25(8)	12(4)	10(3)
	64	Jukes-Cantor	63(53)	52(42)	41(21)	27(17)	25(15)
		Kimura	70(60)	45(35)	22(12)	15(2)	10(0)
		Jin-Nei	75(65)	40(20)	20(10)	16(6)	12(2)

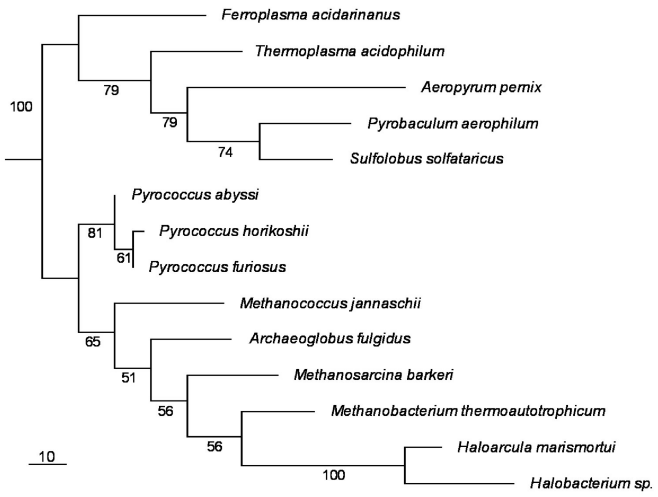


Figure C–9: Maximum likelihood phylogenetic tree for the protein `rpl2e` (89 positions). Numbers close to branches are ML bootstrap scores obtained from the sampled protein sequences using the SeqBoot and Proml (JTT model) programs from the PHYLIP package [85]. Its topology is identical to the tree found by Matte-Taillez *et al* [162, Figure 3].

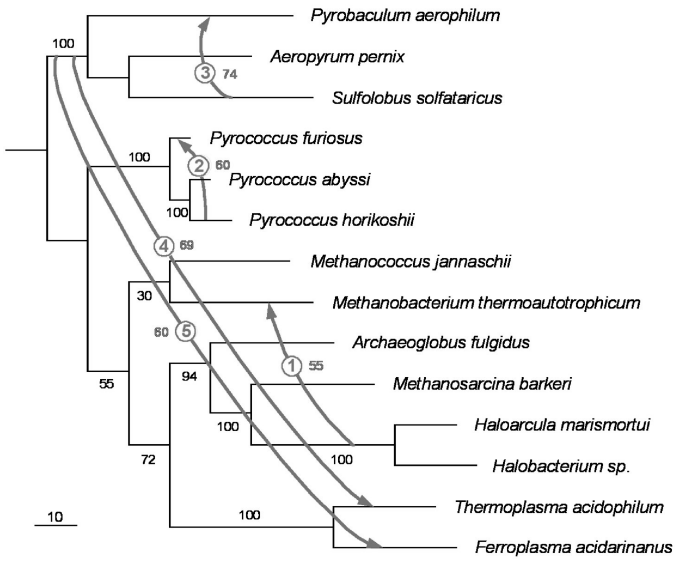


Figure C–10: Species tree (Matte-Taillez *et al.* [162, Fig. 1a], with five reconciliation branches (denoted by arrows). Numbers close to branches are ML bootstrap scores computed by the RELL method upon 2,000 top-ranking trees using the *MOLPHY* program without correction for among-site variation. Numbers on HGT arrows indicate their order of appearance in the unique gene transfer scenario found by the HGT detection method. Bootstrap scores for transfers are indicated by numbers close to arrow circles. Arrows 4 and 5 depict the HGTs between the clades of *Thermoplasmatales* and *Crenarchaeota* also predicted by Matte-Taillez *et al* [162].

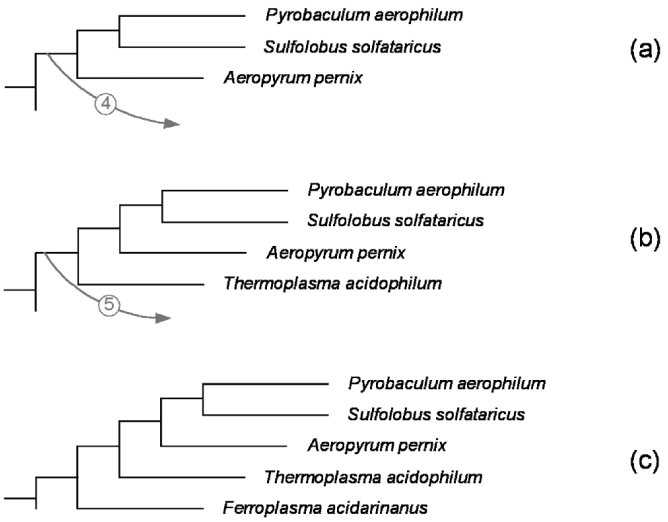


Figure C–11: Changes in the *Crenarchaeota-Thermoplasmatales* cluster occurring after the addition of HGT branches 4 and 5. (a) This cluster after the transfer 3; the species *Thermoplasma acidophilum* joins the *Crenarchaeota* cluster. (b) This cluster after the transfer 4; the species *Ferroplasma acidarmanus* is added to the clade comprising three *Crenarchaeota* and *Thermoplasma acidophilum*. (c) This cluster after the transfer 5.

APPENDIX D
Dynamic programming approach for ancestral Profile-profile alignment

D.1 Preface

This appendix presents only a preliminary dynamic programming algorithmic approach to the ancestral profile alignment. This algorithm will be useful for refining multiple sequence alignment as well as to the joint inference of phylogenetic tree and multiple sequence alignment.

D.2 Importance of profile sequences

Profiles are commonly used in multiple sequence alignment of protein to represent alignment column structure. A profile sequence p of length L indicates for each position of the sequence the probability of each character (A, C, G, T and gap for DNA profiles). The profiles are commonly built using position specific scoring matrix PSSM or HMM profiles [67]. Profiles are involved in alignment in two ways: (1) a sequence is aligned to an existing multiple alignment represented by a profile (sequence-profile alignment) such as PSI-BLAST [2] and HMMAlign from HMMER [69]; (2) two profiles are aligned (profile-profile alignment) such as COACH [73] and MUSCLE [71]. COACH does not perform directly profile-profile alignment, it aligns multiple sequences alignment to a built profile-HMM. Due to complicated recursion relations, it cannot be applied to large data sets. Profile-profile alignment is also the iterated step of ClustalW [112]. Usually the methods of profile-profile alignments are variants of well-known pairwise sequence alignments Needleman-Wunsch for global alignment [180] and Smith-Waterman for local alignment [222]. They differ on the choice of the scoring functions for an

aligned pair of profile (sum over position scores plus affine gap penalties). Here, we define new variant of the global alignment algorithm that produces fast solution. It will use an approach to compute the score according to the presence or absence of characters, that permits to easily handle the problem of scoring affine gap penalties in profiles. A preliminary version of this dynamic programming algorithm is presented here.

Let A and B be two profiles of lenght respectively L_A and L_B such that A is the most recent ancestor of B. Thus $A[i, b]$ is the probability of having the character $b \in \{A, C, G, T, -\}$ in the position $i \in [1..L_A]$.

D.3 Consecutive pair-aligned score.

Let $score(x', y', x, y, i, j)$ be the score of having 2 consecutive aligned columns where $x', y', x, y \in \{0, 1\}$ and x, y are respectively in column i and j. $x'x$ are the 2 consecutive characters of the first sequence and $y'y$ are the consecutive characters of the second sequence. We obtain:

$$score(x', y', x, y, i, j) = \begin{cases} (\sum_{a\in\{A,C,G,T\}} \sum_{b\in\{A,C,G,T\}} (A[i, a][j, b]) \times & \\ Pscore(a, b)) & \text{x=1, y=1} \\ (\sum_{a\in\{A,C,G,T\}} (A[i, a] \times B[j, -]) \times & \\ GapExtensionPenalty(a, -)) & \text{x=1,y=0, x'=1} \\ (\sum_{a\in\{A,C,G,T\}} (A[i, a] \times B[j, -]) \times & \\ GapStartPenalty(a, -)) & \text{x=1,y=0, y'=1} \\ (\sum_{a\in\{A,C,G,T\}} (A[i, -] \times B[j, a]) \times & \\ GapExtensionPenalty(-, a)) & \text{x=0,y=1, x'=0} \\ (\sum_{a\in\{A,C,G,T\}} (A[i, -] \times B[j, a]) \times & \\ GapStartPenalty(-, a)) & \text{x=0,y=1, x'=1,} \\ 0 & otherwise \end{cases}$$

where $Pscore(a, b)$ is log-odds score of aligning the two characters.

Dynamic programming solution.

Let $X_M(i, j, x, y)$ be the score maximal of aligning $A[1..i]$ to $B[1..j]$ whereas $x, y \in \{0, 1\}$. x and y indicate if either gap or existing characters are considered for the profiles A and B respectively, at the end of the prefixes.

Let $X_I(i, j)$ be the maximal score that can be obtained by aligning $A[i]$ to a new gap when prefixes $A[1..i]$ to $B[1..j]$ are considered. We obtain:

$$X_M(i,j,x,y) = \max \begin{cases} \max_{x',y' \in \{0,1\}} \{X_M(i-1, j-1, x', y') + score(x', y', x, y, i, j)\} \\ X_I(i-1, j-1) + score(1, 0, x, y, i, j) \\ X_D(i-1, j-1) + score(0, 1, x, y, i, j) \end{cases}$$

$$X_I(i,j) = \max \begin{cases} \max_{x' \in \{0,1\}} \{X_M(i-1, j-1, x', 1) + score(x', 1, 1, 0, i, j)\} \\ X_I(i-1, j) + score(1, 0, 1, 0, i, j) \end{cases}$$

$$X_D(i,j) = \max \begin{cases} \max_{y' \in \{0,1\}} \{X_M(i-1, j-1, 1, y') + score(1, y', 1, 0, i, j)\} \\ X_D(i, j-1) + score(0, 1, 0, 1, i, j) \end{cases}$$

The two auxiliary matrices are needed to compute affine gap penalties [67]. To find the best alignment score, we will only take $max_{x,y \in \{1,0\}}\{X_M(L_A, L_B, x, y)$. As usual, traceback will be required to find the optimal profile alignment. The dynamic programming showed can be adapted for different gap penalties for insertion and deletion events. To evaluate the accuracy of the obtained alignment, we can use either known aligned sequences to assess the performance of our method on the number of correctly aligned pairs of characters; or simulated evolution of sequences with different sizes according to a fixed phylogenetic tree.

References

[1] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[2] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[3] M. Angulo and A. Carvajal Rodriguez. Evidence of recombination within human alpha-papillomavirus. *Virology Journal*, 4:33, 2007.

[4] A. Antonsson, O. Forslund, H. Ekberg, G. Sterner, and B.G. Hansson. The ubiquity and impressive genomic diversity of human skin papillomaviruses suggest a commensalic nature of these viruses. *Journal of Virology*, 74(24):11636–1164, 2000.

[5] Y. Bao, S. Federhen, D. Leipe, V. Pham, S. Resenchuk, M. Rozanov, R. Tatusov, and T. Tatusova. Ncbi genomes project. *Journal of Virology*, 78(14):7291–7298, 2004.

[6] E. Bapteste and H. Philippe. The potential value of indels as phylogenetic markers: position of trichomonads as a case study. *Molecular Biology and Evolution*, 19(6):972–977, 2002.

[7] J.-P. Barthélemy and A. Guénoche. *Les arbres et les représentations des proximités*. Paris, Masson, 1988.

[8] J.-P. Barthélemy and A. Guénoche. *Trees and proximity representations*. New York, Wiley, 199.

[9] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik L L Sonnhammer, David J Studholme, Corin Yeats, and Sean R Eddy. The Pfam protein families database. *Nucleic Acids Res*, 32(Database issue):D138–D141, Jan 2004.

[10] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick, and D. Haussler. Ultraconserved elements in the human genome. *Science*, 304:1321–1325, 2004.

[11] S.A. Benner. The past as the key to the present: resurrection of ancient proteins from eosinophils. *Proceedings of the National Academy of Science USA*, 99:4760–4761, 2002.

[12] G. Benson. Tandem repeats finder: A program to analyze dna sequences. *Nucleic acids research*, 27:573–580, 1999.

[13] J. Berger, T. Suzuki, K.A. Senti, J. Stubbs, G. Schaffner, and B.J. Dickson. Genetic mappig with snp markers in drosophila. *Nature Genetics*, 29:475–481, 2001.

[14] M. Blanchette. Computation and analysis of genomic multi-sequence alignments. *The Annual Review of Genomics and Human Genetics*, 8:193–213, 2007.

[15] Mathieu Blanchette, Abdoulaye Banire Diallo, Eric Green, Webb Miller, and Haussler David. *Computational Reconstruction of ancestral DNA sequences*, chapter 11, pages 171–184. in Methods in Molecular Biology: Phylogenomics. Springer, New York, 2008.

[16] Mathieu Blanchette, Eric D Green, Webb Miller, and David Haussler. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res*, 14(12):2412–2423, Dec 2004.

[17] Mathieu Blanchette, W James Kent, Cathy Riemer, Laura Elnitski, Arian F A Smit, Krishna M Roskin, Robert Baertsch, Kate Rosenbloom, Hiram Clawson, Eric D Green, David Haussler, and Webb Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research*, 14(4):708–715, Apr 2004.

[18] A. Boc and V. Makarenov. New efficient algorithm for detection of horizontal gene transfer events. In *3rd Workshop on Algorithms in Bioinformatics (Budapest, 2003) (G. Benson and R. Page, eds.) WABI, Lecture Notes in Comput. Sci., Springer Verlag*, pages 190–201, 2003.

[19] F.X. Bosch, M.M. Manos, N. Muñoz, M. Sherman, A.M. Jansen, J. Peto, M.H. Schiffman, V. Moreno, R. Kurman, and K.V. Shan. Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. international biological study on cervical cancer (ibscc) study group. *Journal of the National Cancer Institute*, 87(11):796–802, 1995.

[20] G. Bourque, P.A. Pevzner, and G Tesler. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Research*, 14(4):507–16, 2004.

[21] Guillaume Bourque and Pavel A. Pevzner. Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. *Genome Res.*, 12(1):26–36, 2002.

[22] Guillaume Bourque, Glenn Tesler, and Pavel A Pevzner. The convergence of cytogenetics and rearrangement-based models for ancestral genome reconstruction. *Genome Res*, 16(3):311–313, Mar 2006.

[23] K. Bradley and I Holmes. Transducers: an emerging probabilistic framework for modeling indels on trees. *Bioinformatics*, 23(23):3258–3262, 2007.

[24] N. Bray, I. Dubchak, and L Pachter. MAVID: a global alignment program. *Genome Research*, 13:97–102, 2003.

[25] Nicolas Bray and Lior Pachter. MAVID: constrained ancestral alignment of multiple sequences. *Genome Research*, 14(4):693–699, Apr 2004.

[26] M. Brudno, S. Malde, A. Poliakov, C.B. Do, O. Couronne, I. Dubchak, and S. Batzoglou. Glocal alignment: finding rearrangements during alignment. *Bioinformatics*, 19 Suppl 1.:i54–i62, 2003.

[27] Michael Brudno, Chuong B Do, Gregory M Cooper, Michael F Kim, Eugene Davydov, Eric D Green, Arend Sidow, and Serafim Batzoglou. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research*, 13(4):721–731, Apr 2003.

[28] D. Bryant and V. Moulton. Neighbor-net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology Evolution*, 21:255–265, 2004.

[29] C. Büchen-Osmond. *Taxonomy and Classification of Viruses*, pages 1217–1226. ASM Press, Washington, DC, 8th edition edition, 2003.

[30] J.H. Camin and R.R. Sokal. Pa method for deducing branching sequences in phylogeny. *Evolution*, 19:311–326, 1965.

[31] P.K. Chan, J.L. Cheung, T.H. Cheung, K.W. Lo, S.F. Yim, S.S. Siu, and J.W. Tang. Profile of viral load, integration, and e2 gene disruption of hpv58 in normal cervix and cervical neoplasia. *Journal of Infectious Diseases*, 196(6):868–875, 2007.

[32] S.Y. Chan, H. Delius, A.L. Halpern, and Bernard H.U. Analysis of genomic sequences of 95 papillomavirus types: uniting typing, phylogeny, and taxonomy. *Journal of Virology*, 69(5):3074–3083, 1995.

[33] S.W. Chang, J.A. Ugalde, and M.V. Matz. Applications of ancestral protein reconstruction in understanding protein function: Gfp-like proteins.

Methods in Enzymology, 395:652–670, 2005.

[34] M.A. Charleston. Jungle: a new solution to the host/parasite phylogeny reconciliation problem. *Math. Biosci.*, 149:191–223, 1998.

[35] L. Chindelevitch, Z. Li, E. Blais, and M. Blanchette. On the inference of parsimonious indel evolutionary scenarios. *Journal of Bioinformatics and Computational Biology*, 4(3):721–44, 2006.

[36] L. Choongho and A.L. Laimonis. Role of the pdz domain-binding motif of the oncoprotein e6 in the pathogenesis of human papillomavirus type 31. *Journal of Virology*, 78(22):12366–12377, 2004.

[37] Y.-L Chow and R. Schwartz. The n-best algorithm: An efficient procedure for finding top n sentence hypotheses. In *in Proceedings Speech and Natural Language Workshop October 1989, Cape Cod, Massachusetts,*, pages 199–202, 1989.

[38] P. Cliften, L. Hillier, L. Fulton, T. Graves, T. Miner, W. Gish, R. Waterston, and M. Johnston. Surveying saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. *Genome Research*, 11:1175–1186, 2001.

[39] A.-L. Combita, A. Touzé, L. Bousarghin, N.D. Christensen, and P. Coursaget. Identification of two cross-neutralizing linear epitopes within the l1 major capsid protein of human papillomaviruses. *Journal of Virology*, 76(13):6480–6486, 2002.

[40] The International Hapmap Consortium. A haplotype map of the human genome. *Nature*, 437:1299–1320, 2005.

[41] G.M. Cooper, M. Brudno, E.D. Green, S. Batzoglou, and A. Sidow. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Research*, 13(12):2507–2518, Apr 2003.

[42] G.M. Cooper, E.A. Stone, G. Asimenos, NISC Comparative Sequencing Program, E.D. Green, S. Batzoglou, and A. Sidow. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research*, 15(17):901–913, 2005.

[43] C. Darwin. On the origin of species. *John Murray, London*, 1859.

[44] E. Dawson, Y. Chen, S. Hunt, L.J. Smink, A. Hunt, K. Rice, S. Livingston, S. Bumpstead, R. Bruskiewich, P. Sham, et al. A snp resource for human chromosome 22: Extracting dense clusters of snps from the genomic sequence. *Genome Research*, 11:170–178, 2001.

[45] Nicole de la Chaux, Philipp Messer, and Peter Arndt. Dna indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evolutionary Biology*, 7(1):191, 2007.

[46] E.M. de Villiers, C. Fauquet, T.R. Broker, H.U. Bernard, and H Zur Hausen. Classification of papillomaviruses. *Virology*, 324(1):17–27, 2004.

[47] A. L. Delcher, S. Kasti, R.D. Fleischmann, J. Peterson, and S. L. White, O. Salzberg. Alignment of whole genomes. *Nucleic Acids Research*, 27:2369–2376, 1999.

[48] E. Denamur, G. Lecointre, P. Darlu, et al. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell*, 103:711–721, 2000.

[49] R.C. Deonier, S. Tavaré, and M.S. Waterman. *Computational Genome Analysis*. Springer, New York., 2005.

[50] H. Deveau, M.-C. Labrie, S.J. Chopin, and M. Moineau. iodiversity and classification of lactococcal phages. *Applied and Environmental Microbiology*, pages 4338–4346, 2006.

[51] C. Dewey, J.Q. Wu, S. Cawley, M. Alexandersson, R. Gibbs, and L. Pachter. Accurate identification of novel human genes through simultaneous gene prediction in human, mouse, and rat. *Genome Research*, 14:661–664, 2004.

[52] A.B. Diallo. *Une nouvelle methode efficace pour l´estimation des données manquantes en vue de l´inférence phylogénétique.* Maîtrise en Informatique. Université du Québec à Montréal, 2006.

[53] A.B. Diallo, D. Badescu, V. Makarenkov, and M. Blanchette. An evolutionary study of the human papilloma viruses. *Research in Computational Biology-Comparative Genomics, Paris*, pages 128–140, 2008.

[54] A.B. Diallo, D. Badescu, V. Makarenkov, and M. Blanchette. Phylogeny of the human papilloma viruses and carcinome classification. In *Proceeding of the first joint meeting of the Classification and Data analysis Group and the French Classification Society, Caserta, Italy*, pages 285–288, 2008.

[55] A.B. Diallo, D. Badescu, V. Makarenkov, and M. Blanchette. A whole genome study and identification of specific carcinogenic regions of the human papilloma viruses. *Journal of Computational Biology*, page Submitted, 2009.

[56] A.B. Diallo and E. Gaul. Visualization of the n-best indel likelihood scenarios. *Bioinformatics, Application note*, page In preparation, 2009.

[57] A.B. Diallo, V. Makarenkov, , M. Blanchette, and F-J. Lapointe. A new efficient method for assessing missing nucleotides in dna sequences in the framework of a generic evolutionary model. *Proceedings of the meeting of the International Federation of Classification Societies 2006, Data Science and Classification. eds Batagelj, V., Bock, H.H., Ferligo j, A., Ziberna, A., Springer Verlag, Ljublijana*, pages 333–340, 2006.

[58] A.B. Diallo, V. Makarenkov, and M. Blanchette. Exact and heuristics methods to indel maximum likelihood problem. *Journal Computational Biology*, 14(4):446–461, 2007.

[59] A.B. Diallo, V. Makarenkov, and M. Blanchette. Ancestors 1.0: A web server for ancestral sequence reconstruction. *Bioinformatics, Application note*, page In preparation, 2009.

[60] A.B. Diallo, V. Makarenkov, and M. Blanchette. Evolutionary score for the multiple sequence alignment refinement and the joint inference of phylogenies and multiple sequence alignment. *Bioinformatics*, page In preparation, 2009.

[61] A.B. Diallo, V. Makarenkov, and F-J. Lapointe. A new effective method for estimation of missing values in the sequence data prior to phylogenetic analysis. *Evolutionary Bioinformatics Online*, 2:127–135, 2006.

[62] A.B. Diallo, V. Makarenkov, and Blanchette M. Finding maximum likelihood indel scenarios. In *Comparative Genomics*, pages 171–185. Springer, 2006. LNCS vol. 4205.

[63] A.B. Diallo, D. Nguyen, A. Boc, and V. Makarenkov. étude de la classification des bactériophages. page In preparation, 2009.

[64] Ab. Ba. Diallo, Al. Bo. Diallo, and V. Makarenov. Une nouvelle méthode efficace pour l'estimation des données manquantes en vue de l'inférence phylogénétique. In *Proceeding of the 12th meeting of Société Francophone de Classification*, pages 121–125, 2005.

[65] Chuong B. Do, Mahathi S.P. Mahabhashyam, Michael Brudno, and Serafim Batzoglou. Probcons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*, 15:330–340, 2005.

[66] W.F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284:2124–2129, 1999.

[67] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998.

[68] B. E. Dutilh, M. A. Huynen, W. J. Bruno, and B. Snel. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *Journal of Molecular Evolution*, 58:527–539, 2004.

[69] S. Eddy. Profile hidden markov models. *Bioinformatics*, 14:755–763, 1998.

[70] W. Edelmann, K. Yang, and A.and others Umar. Mutation in the mismatch repair genemsh6 causes cancer susceptibility. *Cell*, 91:476–477, 1996.

[71] Robert C. Edgar. Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 2004.

[72] Robert C. Edgar and Kimmen Sjolander. Satchmo: sequence alignment and tree construction using hidden markov models. *Bioinformatics*, 19(11):1404–1411, 2003.

[73] Robert C. Edgar and Kimmen Sjolander. Coach: Profile-profile alignment of protein families using hidden markov models. *Bioinformatics*, 20(8):1309–1318, 2004.

[74] Robert C. Edgar and Kimmen Sjolander. A comparison of scoring functions for protein sequence profile alignment. *Bioinformatics*, 20(8):1301–1308, 2004.

[75] A. W. F. Edwards and L. L. Cavalli-Sforza. Reconstruction of evolutionary trees. In V. H. Heywood and J. McNeill, editors, *Phenetic and Phylogenetic Classification*, volume 6 of *Systematics Association Publication*, pages 67–76. London, 1964.

[76] E. Eizirik, W.J. Murphy, and S. J. O'Brien. Molecular dating and biogeography of the early placental mammal radiation. *Journal of Heredity*,

92(2):212–219, 2002.

[77] N. El-Mabrouk and F. Lisacek. Very fast identification of RNA motifs in genomic dna. *Journal of Molecular Biology*, 264:46–55, 1996.

[78] W. Enard, M. Przeworski, S.E. Fisher, C.S. Lai, V. Wiebe, T. Kitano, A.P. Monaco, and S. Paabo. Molecular evolution of foxp2, a gene involved in speech and language. *Nature*, 418(6900):869–872, 2002.

[79] ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306:636–640, 2004.

[80] J.S. Farris. Methods for computing wagner trees. *Systematic Zoology*, 19:83–92, 1970.

[81] J.S. Farris. Phylogenetic analysis under dollo's law. *Systematic Zoology*, 26:77–88, 1977.

[82] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[83] J. Felsenstein. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society*, 16:183–196, 1981.

[84] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:738–791, 1985.

[85] J. Felsenstein. Phylip - phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

[86] J. Felsenstein. An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology*, 47:101–111, 1997.

[87] J. Felsenstein. *Inferring phylogenies*. Sinauer Associate, 2003.

[88] J. Felsenstein and G. Churchill. A hidden markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13:93–104, 1996.

[89] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.

[90] Walter M. Fitch. Toward defining the course of evolution: Minimum change for a specified tree topology. *Systematic Zoology*, 20:406–416, 1971.

[91] J. Fredslund, J. Hein, and T. Scharling. A large version of the small parsimony problem. In *Proceedings of the 4th Workshop on Algorithms in Bioinformatics (WABI)*, 2004.

[92] O. Gascuel. Concerning the nj algorithm and its unweighted version, unj. *Mathematical hierarchies and Bit Biology*, 45:363–374, 1997.

[93] O. Gascuel. An improved version of nj algorithm based on a simple model of sequence data. *Molecular Biology Evolution*, 14:685–695, 1997.

[94] E.A. Gaucher, M.F. Thomson, M.F. Burgan, and S.A. Benner. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature*, 425:285–288, 2003.

[95] C. Gaudreau and A.B. Diallo. A computational approach for the classification of $\gamma-$proteobacteria family. *Molecular Biology and Evolution*, page to be submitted, 2009.

[96] Genbank. `http://www.ncbi.nlm.nih.gov/Genbank`, 2000.

[97] G. Glazko, A. Gordon, and A. Mushegian. The choice of optimal distance measure in genome-wide datasets. *Bioinformatics*, pages iii3–iii11, 2005.

[98] M. Goodman, J. Barnabas, G. Matsuda, and G.W. Moore. Molecular evolution in the descent of man. *Nature*, 233:604–613, 1971.

[99] D.A. Graham and C.S. Herrington. Hpv-16 e2 gene disruption and sequence variation in cin 3 lesions and invasive squamous cell carcinomas of the cervix: relation to numerical chromosome abnormalities. *Molecular Pathology*, 53:201–206, 2000.

[100] S. Guindon and O. Gascuel. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Molecular Biology Evolution*, 19:534–543, 2002.

[101] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5):696–704, 2003.

[102] S.K. Gupta, J.D. Kececioglu, and A.A. Schaffer. Improving the pratical space and time efficiency of the shortest-paths approach to sum-of-pairs multiple sequence alignment. *Journal of Computational Biology*, 2:459–472, 1995.

[103] M. Hallet and J. Lagergren. Efficient algorithms for lateral gene transfer problems. *RECOMB (N. El-Mabrouk, T. Lengauer, and D. Sankoff, eds.), ACM, New York*, pages 149–156, 2001.

[104] M. Hallet, J. Lagergren, and A. Tofigh. Simultaneous identification of duplications and lateral transfers. *RECOMB (N. El-Mabrouk, T. Lengauer, and D. Sankoff, eds.), ACM, New York*, pages 347–356, 2004.

[105] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174, 1985.

[106] Bernard Haubold and Thomas Wiehe. *Introduction to Computational Biology*. Birkhauser Verlag, Boston-Berlin., 2006.

[107] J. Hein. A method that simultaneously aligns, finds the phylogeny and reconstructs ancestral sequences for any number of ancestral sequences. *Molecular Biology and Evolution*, 6(6):649–668, 1989.

[108] J. Hein. An algorithm for statistical alignment of sequences related by a binary tree. *Pac. Symp. Biocomp, world scientific*, pages 179–190, 2001.

[109] J. Hein, J.L. Jensen, and C.N.S. Pedersen. Recursions for statistical multiple alignment. *Proceedings of the National Academy of Science USA*, 100(6):4960–4965, 2003.

[110] J. Hein, T. Jiang, L. Wang, and K. Zhang. On the complexity of comparing evolutionary trees. *Discrete Appl. Math.*, 71:153–169, 1996.

[111] R.W. Hendrix. Bacteriophages: evolution of the majority. *Theoretical Population Biology*, 61:471–480, 2002.

[112] D.G. Higgins, Thompson J.D., and T.J. Gibson. Using CLUSTAL for multiple sequence alignments. In Russell F. Doolittle, editor, *Computer Methods for Macromolecular Sequence Analysis*, volume 266 of *Methods in Enzymology*, pages 383–401. Academic Press, New York, 1996.

[113] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of American Statistical Association*, 58:13–27, 1963.

[114] I. Hofacker, W. Fontana, P Stadler, S. Bonhoeffer, S. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[115] I. Holmes and W.J. Bruno. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17(9):803–820, 2001.

[116] A. Hudek and D.G. Brown. Ancestral sequence alignment under optimal conditions. *BMC Bioinformatics*, 6:273:1–14, 2005.

[117] J.P. Huelsenbeck. When are fossils better than existent taxa in phylogenetic analysis. *Systematic Zoology*, 40:458–469, 1991.

[118] J.P. Huelsenbeck and F. Ronquist. MrBayes: Bayesian inference of phylogeny. *Bioinformatics*, 17:754–755, 2001.

[119] J.P. Huelsenbeck, F. Ronquist, R. Nielsen, and J.P. Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294:2310–2314, 2001.

[120] L. Hufford. Rosidaea and their relationships to other nonmagnoliid dicotyledons: A phylogenetic analysis using morphological and chemical data. *Annals of the Missouri Botanical Garden*, 79:218–248, 1992.

[121] D.H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology Evolution*, 23:254–267, 2006.

[122] International Human Genome Sequencing Consortium, E. Lander, et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[123] A.W. Jarvis, G.F. Fitzgerald, M. Mata, A. Mercenier, H. Neve, I. Powell, C. Ronda, M. Saxelin, and M. Teuber. Species and type phages of lactococcal bacteriophages. *Intervirology*, 32:2–9, 1991.

[124] T.M. Jermann, J.G. Optiz, J. Stackhouse, and S.A. Benner. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature*, 374:57–59, 1995.

[125] L. Jin and M. Nei. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Molecular Biology Evolution*, 7:82–102, 1990.

[126] T.H. Jukes and C.R. Cantor. *Evolution of protein molecules*, pages 21–123. in H. N. Munro, ed., New York, 1969.

[127] J. Jurka. Repbase update: a database and an electronic journal of repetitive elements. *Trends on Genetics*, 16(19):418–420, 2002.

[128] D Karolchick et al. The ucsc genome browser database. *Nucleic Acids Research*, 31:51–54, 2003. `http://genome.ucsc.edu`.

[129] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30:3059–3066, 2002.

[130] W. James Kent, Robert Baertsch, Angie Hinrichs, Webb Miller, and David Haussler. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A*, 100(20):11484–11489, Sep 2003.

[131] J. Kim and S Sinha. Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. *Bioinformatics*, 23(3):289–297, 2007.

[132] M. Kimura. A simple model for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[133] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, 29:170–179, 1989.

[134] C. Korostensky, Stege U., and G.H. Gonnet. Algorithms for improving multiple sequence alignments and building evolutionary trees. *RECOMB*, 2000.

[135] N.M. Krishnan, H. Seligman, C. Stewart, A.P. Jason de Koning, and D.D. Pollock. Ancestral sequence reconstruction in primate mitochondrial dna: Compositional bias and effect on functional inference. *Molecular Biology and Evolution*, 21(10):1871–1883, 2004.

[136] A. Krogh, M. Brown, S. Mian, M. Sjölander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. Technical Report UCSC-CRL-93-32, Department of Computer and Information Sciences, University of California at Santa Cruz, 1993.

[137] D. Kulp, D. Haussler, M.G. Reese, and F.H. Eeckman. A generalized hidden Markov model for the recognition of human genes in DNA. In

D. J. States, P. Agarwal, T. Gaasterland, L. Hunter, and R. Smith, editors, *Proc. Conf. On Intelligent Systems in Molecular Biology '96*, pages 134–142. AAAI/MIT Press, 1996. St. Louis, Mo.

[138] S. Kumar, J. Dudley, M. Nei., and K. Tamura. Mega: A biologist-centric software for evolutionary analysis of dna and protein sequences. *Briefings in Bioinformatics*, 9:299–306, 2008.

[139] M. Kunher and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology Evolution*, 11:459–468, 1996.

[140] B. Larget, D. Simon, J.B. Kadane, and D. Sweet. A bayesian analysis of metazoan mitochondrial genome arrangements. *Molecular Biology and Evolution*, 22:486–495, 2005.

[141] J.G. Lawrence and H. Ochman. Amelioration of bacterial genomes: rates of change and exchange. *Journal of Molecular Evolution*, 44:383–397, 1997.

[142] L. Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York, 1986.

[143] C. Lee. Generating consensus sequences from partial order multiple sequence alignment graphs. *Bioinformatics*, 19:999–1008, 2003.

[144] P. Legendre and V. Makarenkov. Reconstruction of biogeographic and evolutionary networks using reticulograms. *Systematic Biology*, 51:199–216, 2002.

[145] P. (guest ed.) Legendre. Special section on reticulate evolution. *Journal of Classification*, 17:153–195, 2000.

[146] A. Lemmon and M. Milinkovitch. The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. *Proceedings of the National Academy of Science USA*, 99:10516–10521, 2002.

[147] W-H. Li. *Molecular evolution*. Sunderland, Massachusetts: Sinauer Associate, 1997.

[148] J. Liu, G. Glazko, and A. Mushegian. Protein repertoire of double-stranded dna bacteriophages. *Virus Research*, 117:68–80, 2006.

[149] Gabriela G Loots and Ivan Ovcharenko. rVISTA 2.0: evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue):217–221, Jul 2004.

[150] A. Loytynoja and N. Goldman. An algorythm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Science USA*, 102(38):10557–10562, 2005.

[151] G. Lunter, C. Ponting, and J. Hein. Genome-wide identification of human functional dna using a neutral indel model. *Genome Research*, 2(1):e5, 2006.

[152] G.A. Lunter, I. Miklos, Y.S. Song, and J. Hein. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J Computational Biology*, 10(6):869–89, 2003.

[153] D.R. Maddison and K.-S. Schulz. The tree of life web project. `http://tolweb.org`, 2004.

[154] W.P. Maddison. Gene trees in species trees. *Systematic Biology*, 46:523–536, 1997.

[155] V. Makarenkov. T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. *Systematic Biology*, 17:664–668, 2001.

[156] V. Makarenkov, A. Boc, Diallo Al. Bo., and Diallo Ab.Ba. Algorithms for detecting complete and partial horizontal gene transfers: Theory and practice. In *Data Mining and Mathematical Programming, P.M. Pardalos and P. Hansen eds., CRM Proceedings and AMS Lecture Notes*, volume 45, pages 159–179, 2008.

[157] V. Makarenkov, D. Kevorkov, and P. Legendre. *Phylogenetic Network Reconstruction Approaches*, chapter 3, pages 61–97. International Elsevier Series., 2006.

[158] V. Makarenkov and F-J. Lapointe. A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics*, 20:2113–2121, 2004.

[159] V. Makarenkov and B. Leclerc. An algorithm for the fitting of a phylogenetic tree according to a weighted least-squares criterion. *Journal of Classification*, 16:3–26, 1999.

[160] V. Makarenov, A. Boc, F. Delwiche, A.B. Diallo, and H. Philippe. New efficient algorithm for modeling partial and complete gene transfer scenarios. In *Data Science and Classification (V. Batagelj, H.-H. Bock, A. Ferligoj, and A. Ziberna, eds.), IFCS 2006, Studies in Classification, Data Analysis, and Knowledge Organization, part VIII, Springer Verlag*, pages 341–349, 2006.

[161] Elliott H Margulies, Mathieu Blanchette, David Haussler, and Eric D Green. Identification and characterization of multi-species conserved sequences. *Genome Research*, 13(12):2507–2518, Dec 2003.

[162] O. Matte-Tailliez, C. Brochier, P. Forterre, and H. Philippe. Archaeal phylogeny based on ribosomal proteins. *Molecular Biology Evolution*, 19:631–639, 2002.

[163] I. Miklos, A. Lunter, and I. Holmes. A long indel model for evolutionary sequence alignment. *Molecular Biology and Evolution*, 21(3):529–540, 2004.

[164] W. Miller. Personal communication. 2006.

[165] R.E. Mills, C.T. Luttig, C.E. Larkins, A. Beauchamp, C. Tsui, W.S. Pittard, and S.E. Devine. An initial map of insertion and deletion (indel) variation in the human genome. *Genome Research*, 16:1182–1190, 2006.

[166] B. Mirkin. Mapping gene family data onto evolutionary trees. In *Comptes rendus des 11es Rencontres de la Sociéte Francophone de Classification, (M. Chavent, O. Dordan, C. Lacomblez, M. Langlais, and B. Patouille, eds.), University of Bordeaux*, pages 61–68, 2004.

[167] B. Mirkin, T.I. Fenner, M. Galperin, and E. Koonin. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evolutionary Biology*, 3:2, 2003.

[168] B. Mirkin and E. Koonin. A top-down method for building genome classification trees with linear binary hierarchies. *DIMACS series in Discrete Mathematics Theoretical Computer Science*, 2003.

[169] B.M.E. Moret, L. Nakhkeh, T. Warnow, C.R. Linder, A. Tholse, A. Padolina, J. Sun, and R. Timme. Phylogenetic networks: modeling, reconstructibility, and accuracy. *IEEE/ACM Trans. on Comput. Biol. and Bioinf*, 1:13–23, 2004.

[170] B. Morgenstern, K. Frech, A. Dress, and T. Werner. DIALIGN: Finding local similarities by multiple sequence alignment. *Bioinformatics*, 14:290–294, 1998.

[171] Burkhard Morgenstern. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15(3):211–218, 1999.

[172] A.M. Moses, D.Y. Chiang, D.A. Pollard, V.N. Iyer, and M.B. Eisen. Monkey: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biology*, page R98, 2004.

[173] William J Murphy, Guillaume Bourque, Glenn Tesler, Pavel Pevzner, and Stephen J O'Brien. Reconstructing the genomic architecture of mammalian ancestors using multispecies comparative maps. *Hum Genomics*, 1(1):30–40, Nov 2003.

[174] W.J. Murphy, E. Eizirik, W. 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.

[175] N. Muñoz. Human papillomavirus and cancer: the epidemiological evidence. *Journal of Clinical Virology*, 19(1-2):1–5, 2000.

[176] N. Muñoz, F.X. Bosch, X. Castellsagué, M. Daz, S. de Sanjose, D. Hammouda, K.V. Shah, and C.J. Meijer. Against which human papillomavirus types shall we vaccinate and screen? the international perspective. *International Journal of Cancer*, 111:278–285, 2004.

[177] N. Muñoz, F.X. Bosch, S. de Sanjosé, R. Herrero, X. Castellsagué, K.V. Shah, P.J.F. Snijders, and C.J.L.M. Meijer. Epidemiologic classification of human papillomavirus types associated with cervical cancer. *New England Journal of Medecine*, 384:518–527, 2003.

[178] L. Nakhleh, D. Ruths, and H. Innan. *Gene trees, species trees, and species networks*, pages 1–27. in R. Guerra and D. Allison, eds., 2005.

[179] A. Narechania, Z. Chen, R. DeSalle, and R.D. Burk. Phylogenetic incongruence among oncogenic genital alpha human papillomaviruses. *Journal of Virology*, 79:15503–15510, 2005.

[180] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[181] D. Nguyen, A. Boc, A.B. Diallo, and V. Makarenkov. Classification des bactériophages. *Proceedings of the 13th meeting of the Société Francophone de Classification*, pages 161–165, 2007.

[182] C. Notredame and D.G. Higgins. Saga: sequence alignment by genetic algorithm. *Nucleic Acids Research*, 24:1515–1524, 1996.

[183] C. Notredame, E.A. O'Brien, and D.G. Higgins. Raga: RNA sequence alignment by genetic algorithm. *Nucleic Acids Research*, 25:4570–80, 1997.

[184] G. Osen, H. Matsuda, R. Hagstrom, and R. Overbeek. Fastdnaml: A tool for construction of phylogenetic trees of dna sequences using maximum likelihood. *Computational Application in Bioscience*, 10:41–48, 1994.

[185] S. Ota and W.-H. Li. A hybrid algorithm for the neighbor-joining and maximum-likelihood methods. *Molecular Biology and Evolution*, 18:1983–1992, 2001.

[186] Cordano P., V. Gillan, S. Bratlie, V. Bouvard, L. Banks, M. Tommasino, and M.S. Campo. The e6e7 oncoproteins of cutaneous human papillomavirus type 38 interfere with the interferon pathway. *Virology*, 377(2):408–418, 2008.

[187] R.D.M. Page. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43:58–77, 1994.

[188] R.D.M. Page and A. Charleston. From gene to organismal phylogeny: reconciled trees. *Bioinformatics*, 14:819–820, 1998.

[189] R.D.M. Page and A. Charleston. Trees within trees: phylogeny and historical associations. *Trends on Ecology and Evolution*, 13:356–359, 1998.

[190] L. Patthy. *Protein Evolution*. Blackwell, Oxford, 2nd edition, 2007.

[191] L. Pauling and E. Zuckerkandl. Molecular 'restoration studies' of extinct forms of life. *Acta chemica Scandinavica*, 17:9–16, 1963.

[192] H.C. Prasanna and R. Mathura. Detection and frequency of recombination in tomato-infecting begomoviruses of south and southeast asia. *Virology journal*, 4:111:doi:10.1186/1743–422X–4–111, 2007.

[193] J.L. Prétet, J.F. Charlot, and C. Mougin. Virological and carcinogenic aspects of hpv. *Bulletin Academic National de Medecine*, 191(3):611–613, 2007.

[194] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[195] A. Rambault and N. Grassly. Seqgen: An application for the monte carlo simulation of dna sequences evolution along phylogenetic trees. *Bioinformatics*, 13:235–238, 2004.

[196] B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, 43:304–311, 1996.

[197] Rat Genome Sequencing Project Consortium, Gibbs, et al. Genome sequence of the brown norway rat yields insights into mammalian evolution. *Nature*, 428:493–521, 2004.

[198] V. Rawnez and Gascuel O. Improvement of distance-based phylogenetic methods by a local maximum likelihood approach using triplets. *Molecular Biology and Evolution*, 19:1952–1963, 2002.

[199] S. Ricagno, V. Campanacci, S. Blangy, S. Spinelli, D. Tremblay, S. Moineau, M. Tegoni, and C. Cambillau. Crystal structure of the receptor-binding protein head domain from l.lactis phage bil170. *Journal of Virology*, pages 9331–9335, 2006.

[200] E. Rivas. Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinformatics*, 6(1):63, 2005.

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

[202] F.J. Rohlf. Phylogenetic models and reticulations. *Journal of classification*, 17(2):185–189, 2000.

[203] F. Rohwer and R. Edwards. The phage proteomic tree: a genome-based taxonomy for phage. *Journal of Bacteriology*, 184:4529–4535, 2002.

[204] A. Rokas and P.W.H. Holland. Rare genomic changes as a tool for phylogenetics. *Trends on Ecology*, 15:454–459, 2000.

[205] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology Evolution*, 4:406–425, 1987.

[206] L. Salter and D. Pearl. Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Systematic Biology*, 50:7–17, 2001.

[207] Albin Sandelin, Wyeth W Wasserman, and Boris Lenhard. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, 32(Web Server issue):249–252, Jul 2004.

[208] M.J. Sanderson, A. Purvis, and C. Henze. Phylogenetic supertrees: Assembling the tree of life. *Trends on Ecology and Evolution*, 13:105–109, 1998.

[209] D. Sankoff and R.J. Cedergren. Simultaneous comparison of three or more sequences related by a tree. *In Time warps, string edits and macromolecules: the theory and practice of sequence comparison*, ed. DaK Sankoff:253–263, 1983.

[210] D. D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, 28:35–42, 1975.

[211] S. Sattah and A. Tversky. Phylogenetic similarity trees. *Psychometrika*, 42:319–345, 1977.

[212] S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, and W. Miller. Human-mouse alignments with blastz. *Genome Research*, 13(1):103–7, 2003.

[213] H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16:1114–1116, 1999.

[214] H. Shimodaira and M. Hasegawa. Consel: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17:1246–1247, 2001.

[215] A. Siepel and D. Haussler. Combining phylogenetic and hidden markov models in biosequence analysis. *J Comput Biology*, 11(2-3):413–28, 2004.

[216] A. Siepel, K.S. Pollard, and D. Haussler. New methods for detecting lineage-specific selection. In *Proceedings of the 10th International Conference on Research in Computational Molecular Biology (RECOMB 2006)*, pages 190–205, 2006.

[217] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, Ladeana W

Hillier, Stephen Richards, George M Weinstock, Richard K Wilson, Richard A Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, Aug 2005.

[218] D. Simmon and B. Larget. Bayesian analysis in molecular biology and evolution (bambe), version 2.03beta. *Departement of mathematics and computer science, Duquesne University, Pittsburrgh, Pensilvannia*, 2000.

[219] A. Smit and P. Green. Repeatmasker.

[220] A.F. Smit. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Current Opinion in Genetics Development*, 9:657–663, 1999.

[221] J.F. Smith. Tribal relationships within gesneriaceae: A cladistic analysis of morphological data. *Systematic Botanic*, 21:497–513, 1997.

[222] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, March 1981.

[223] S. Spinelli, V. Campanacci, S. Blangny, S. Moineau, M. Tegoni, and C. Cambillau. Modular structure of the receptor binding proteins of lactococcus lactis phages: the rbp structure of the temperature phages tp901-1. *Journal of Biological Chemistry*, 20:4256–4262, 2006.

[224] S. Spinelli, A. Desmyter, C.T. Verrips, H.J. De Haard, S. Moineau, and C. Cambillau. Lactococcal bacteriophage p2 receptor-binding protein structure suggests a common ancestor gene with bacterial and mammalian viruses. *Nature Structural and Molecular Biology*, 13:85–89, 2005.

[225] M.S. Springer, W.J. Murphy, E. Eizirik, and S.J. O'Brien. Placental mammal diversification and the cretaceous-tertiary boundary. *PNAS*, 100(3):1056–1060, 2003.

[226] J. Stoye, D. Evers, and F. Meyer. Rose: generating sequence families. *Bioinformatics*, 14:2:157–163, 1998.

[227] K. Strimmer and A. Von Haeseler. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 13:964–969, 1996.

[228] D. L. Swofford and G. J. Olsen. Phylogeny reconstruction. In D. M. Hillis and C. Moritz, editors, *Molecular Systematics*, chapter 11, pages 411–501. Sinauer Associates, Sunderland, Massachusetts, 1996.

[229] D.L. Swofford. Paup*. phylogenetic analysis using parsimony (*and other methods). version 4. *Sinauer Associates, Sunderland, Massachusetts*, 2001.

[230] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Molecular Biology and Evolution*, 10:512–526, 1993.

[231] Jijun Tang and Bernard M E Moret. Scaling up accurate phylogenetic reconstruction from gene-order data. *Bioinformatics*, 19 Suppl 1:i305–i312, 2003.

[232] M. Tedrano, D.N. Cooper, J. Stuhrmann, M.and Christodoulou, N.A. Chuzhanova, F. Roudot-Thoraval, P.Y. Boelle, J. Elion, M. Jeanpierre, J. Feingold, R. Couderc, and M. Bahuau. Origin of the prevalent sftpb indel g.1549c ¿gaa (121ins2) mutation causing surfactant protein b (sp-b) deficiency. *American Journal of Medical Genetics*, 140A:62–66, 2006.

[233] The International Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, 2002.

[234] J W Thomas, J W Touchman, R W Blakesley, G G Bouffard, S M Beckstrom-Sternberg, E H Margulies, M Blanchette, A C Siepel, P J

Thomas, J C McDowell, B Maskeri, N F Hansen, M S Schwartz, R J Weber, W J Kent, D Karolchik, T C Bruen, R Bevan, D J Cutler, S Schwartz, L Elnitski, J R Idol, A B Prasad, S-Q Lee-Lin, V V B Maduro, T J Summers, M E Portnoy, N L Dietrich, N Akhter, K Ayele, B Benjamin, K Cariaga, C P Brinkley, S Y Brooks, S Granite, X Guan, J Gupta, P Haghighi, S-L Ho, M C Huang, E Karlins, P L Laric, R Legaspi, M J Lim, Q L Maduro, C A Masiello, S D Mastrian, J C McCloskey, R Pearson, S Stantripop, E E Tiongson, J T Tran, C Tsurgeon, J L Vogt, M A Walker, K D Wetherby, L S Wiggins, A C Young, L-H Zhang, K Osoegawa, B Zhu, B Zhao, C L Shu, P J De Jong, C E Lawrence, A F Smit, A Chakravarti, D Haussler, P Green, W Miller, and E D Green. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424(6950):788–793, Aug 2003.

[235] J.D. Thompson, D.G. Higgins, and T.J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4680, 1994.

[236] J. L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol*, 33(2):114–124, Aug 1991.

[237] J.L. Thorne, H. Kishino, and J. Felsenstein. Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.*, 34:3–16, 1992.

[238] B. Tillman. `http://commons.wikimedia.org/wiki/Image:Tree_of_life.svg`, 2007.

[239] K. Tohru, Atsuro H., F. Masatoshi, Yasuyuki H., Tetsu A., and Masahide I. Binding of high-risk human papillomavirus e6 oncoproteins to the

human homologue of the drosophila discs large tumor suppressor protein. *pnas*, 94(21):11612–11616, 1997.

[240] H. Tseng. Complementary oligonucleotides and the origin of the mammalian involucrin gene. *Gene*, 194:87–95, 1997.

[241] C. Tsui, L.E. Coleman, J.L. Griffiths, E.A. Bennett, S.G. Goodson, J.D. Scott, W.S. Pittard, and S.E. Devine. Single nucleotide polymorphisms (snps) that map to gaps in the human snp map. *Nucleic Acids Research*, 31:4910–4916, 2003.

[242] M. Van Ranst, J.B. Kaplanlt, and R.D. Burk. Phylogenetic classification of human papillomaviruses: Correlation with clinical manifestations. *Journal of General Virology*, 73:2653–2660, 1992.

[243] A. Varsani, E. Van der Walt, L. Heath, E.P. Rybicki, A.L. Williamson, and D.P. Martin. Evidence of ancient papillomavirus recombination. *Journal of General Virology*, 87:2527–2531, 2006.

[244] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[245] A. Von Haeseler and G.A. Churchill. Network models for sequence evolution. *Journal of Molecular Evolution*, 37:77–85, 1993.

[246] Iain M. Wallace, O'Sullivan Orla, and Desmond G. Higgins. Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics*, 21:I408–I414, 2005.

[247] J.T. Wang, L. Ding, E.S. Gao, and Y.Y. Cheng. Analysis on the expression of human papillomavirus type 16 e2 and e6 oncogenes and disruption of e2 in cervical cancer. *Zhonghua Liu Xing Bing Xue Za Zhi*, 28(10):968–971, 2007.

[248] Y. Wang and K.B. Li. An adaptative and iterative algorithm for refining multiple sequence alignment. *Comput. Biol. Chem.*, 28:141–148, 2004.

[249] R.D. Wells. Molecular basis of genetic instability of triplet repeats. *Journal of Biological Chemistry*, 271:2875–2878, 1996.

[250] S.R. Wicks, R.T. Yeh, W.R. Gish, R.H. Waterston, and R.H.A. Plasterk. Rapid gene mapping in caenorhabiditis elegans using a high density polymorphism map. *Nature Genetics*, 28:160–164, 2001.

[251] J.J. Wiens. Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology*, 47:528–538, 1998.

[252] J.J. Wiens. Does adding characters with missing data increase or decrease phylogenetic accuracy. *Systematic Biology*, 52:625–640, 2003.

[253] R. Wilson, G.B. Ryan, G.L. Knight, L.A. Laimins, and S. Roberts. The full-length e1^e4 protein of human papillomavirus type 18 modulates differentiation-dependent viral dna amplification and late gene expression. *Virology*, 362(2):453–460, 2007.

[254] C.R. Woese, R. Gutell, R. Gupta, and H.F. Noller. Detailed analysis of the higher-order of 16s-like ribosomal ribonucleic acids. *Microbiology Review*, 47:621–669, 1983.

[255] X. Xia and Z. Xie. Dambe: Data analysis in molecular biology and evolution. *Journal of Heredity*, 92:371–373, 2001.

[256] Z. Yang. Among-site rate variation and its impact on phylogenetic analysis. *Trends in Ecology and Evolution*, 11(9):367–372, 1996.

[257] Z. Yang. Maximu likelihood models for combined analyses of multiple sequence data. *J of molecular Evolution*, 42:587–596, 1996.

[258] Z. Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *CABIOS*, 13:555–556, 1997.

[259] Z. Yang, S. Kumar, and M. Nei. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141:1641–1650, 1995.

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 14, Number 4, 2007

Pp. 446–461
DOI: 10/1089/cmb.2007.A006

Exact and Heuristic Algorithms for the Indel Maximum Likelihood Problem

ABDOULAYE BANIRE DIALLO,[1,2] VLADIMIR MAKARENKOV,[2]
and MATHIEU BLANCHETTE[1]

ABSTRACT

Given a multiple alignment of orthologous DNA sequences and a phylogenetic tree for these sequences, we investigate the problem of reconstructing the most likely scenario of insertions and deletions capable of explaining the gaps observed in the alignment. This problem, that we called the Indel Maximum Likelihood Problem (IMLP), is an important step toward the reconstruction of ancestral genomics sequences, and is important for studying evolutionary processes, genome function, adaptation and convergence. We solve the IMLP using a new type of tree hidden Markov model whose states correspond to single-base evolutionary scenarios and where transitions model dependencies between neighboring columns. The standard Viterbi and Forward-backward algorithms are optimized to produce the most likely ancestral reconstruction and to compute the level of confidence associated to specific regions of the reconstruction. A heuristic is presented to make the method practical for large data sets, while retaining an extremely high degree of accuracy. The methods are illustrated on a 1-Mb alignment of the CFTR regions from 12 mammals.

Key words: ancestral genome reconstruction, ancestral mammalian genomes, indel maximum likelihood problem, insertions and deletions, tree-HMM.

1. INTRODUCTION

IT HAS RECENTLY BEEN SHOWN that the phylogeny of eutherian mammals is such that an accurate reconstruction of the genome of an early ancestral mammal is possible (Blanchette et al., 2004a). This accurate reconstruction will help on various studies such as adaptation, behavioral changes, and functional divergences (Krishnan et al., 2004). It is also at the core of experimental paleo-molecular biochemistry where sequences of extant taxa are used to predict and resurrect the sequences and functions of ancestral macromolecules (Benner, 2002; Gaucher et al., 2003; Pauling and Zuckerkandl, 1963). The ancestral genome reconstruction procedure involves several difficult steps, including the identification of orthologous regions in different extant species, ordering of syntenic blocks, multiple alignment of orthologous sequences within each syntenic block, and reconstruction of ancestral sequences for each aligned block. This last

[1]McGill Centre for Bioinformatics and School of Computer Science, McGill University, Montréal, Québec, Canada.
[2]Département d'Informatique, Université du Québec à Montréal, Montréal, Québec, Canada.

step involves the inference of the set of substitutions, insertions, and deletions that may have produced a given set of multiply-aligned extant sequences. While the problem of reconstructing substitutions scenarios has been well studied (Fitch, 1971; Felsenstein, 1981), the inference of insertions and deletions scenarios has received less attention (Thorne et al., 1991). Indel evolutionary scenarios are useful for several other problems such as annotating functional regions of extant genomes, including protein-coding regions (Siepel and Haussler, 2004), RNA genes (Rivas, 2005), and other types of functional regions (Siepel et al., 2005). The difficulty of the problem is due in large part to the fact that insertions and deletions (indels) often affect several consecutive nucleotides, so the columns of the alignment cannot be treated independently, as opposed to the maximum likelihood problem for substitutions (Felsenstein, 1981). The reconstruction of the most parsimonious scenario of indels required to explain a given multiple sequence alignment has been shown to be NP-Complete (Chindelevitch et al., 2006) but good heuristics have been developed (Blanchette et al., 2004a; Chindelevitch et al., 2006; Fredslund et al., 2004).

A maximum likelihood reconstruction would be preferable to a most parsimonious reconstruction because it would provide a way of weighing insertions and deletions of various lengths against each other. Moreover, provided an accurate probabilistic model is used, the reconstruction would be more accurate and would allow to estimate the uncertainty to each of its parts. Similarly to statistical alignment approaches (Lunter et al., 2003), which unfortunately remain too slow for genome-wide reconstructions, we seek to gain a richer insight into ancestral sequences and evolutionary processes. In this paper, we thus focus on the problem we call the *Indel Maximum Likelihood Problem (IMLP)*. It consists of inferring the set of insertions and deletions that has the maximal likelihood, according to some fixed evolutionary parameters, and that could explain the gaps observed in a given multiple alignment. An example of the input and output of this problem is shown in Figure 1. Kim and Sinha (2007) have recently proposed an algorithm for a similar problem, although the range of scenarios handled by their Indelign program is limited to non-overlapping indels.

We emphasize that the problem addressed here assumes that the phylogenetic tree and multiple sequence alignment given as input are correct. The robustness of indel scenarios with respect to alignment and tree accuracy has been previously discussed (Blanchette et al., 2004a). The more general problem where the alignment is not given as input but has to be found simultaneously with the ancestral sequences (Hein, 1989) is clearly of great interest but appears significantly more difficult and is not addressed here. We refer the reader elsewhere (Kim and Sinha, 2007; Bray and Pachter, 2004) for interesting first steps in that direction.

Here, we start by giving a formal definition of the Indel Maximum Likelihood Problem. To solve the problem, we use a special type of tree hidden Markov model (tree-HMM), which is a combination of a standard hidden Markov model and a phylogenetic tree. We show how the most likely path through the tree-HMM leads to the most likely indel scenario and how a variant of the standard Viterbi algorithm can solve the problem. Although the size of the HMM is exponential in the number of extant species considered, we show how the knowledge given by the phylogenetic tree and the aligned sequences allows the state space of the HMM to be considerably reduced, resulting in a practical, yet exact, algorithm. We

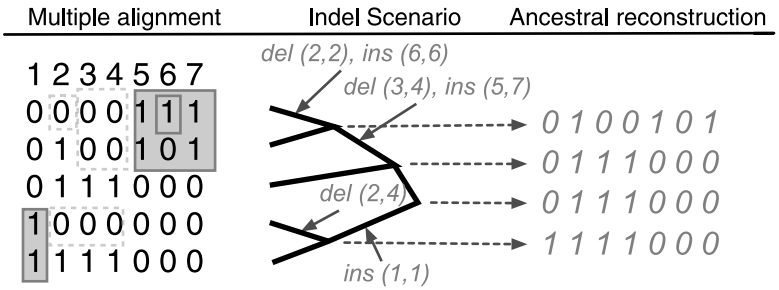


FIG. 1. Example of an input and output to the Indel Maximum Likelihood Problem. The input (in black) consists of the multiple alignment (shown on the left in binary format) and the topology and branch lengths of the phylogenetic tree. The output (in gray and italics) consists of a set of insertions and deletions, placed along the edges of the tree, explaining the gaps (zeros) in the alignment. The dashed (resp. shaded) boxes in the alignment indicate the deletions (resp. insertions) of the scenario shown on the right. This set of operations yields the ancestral reconstruction shown on the right.

also present a heuristic algorithm that almost always gives the right solution and can compute the most likely indels scenarios for more than 20 taxa. Thus, our implementations are able to solve large problems on a simple desktop computer and allow for an easy parallelization. Finally, we assess the complexities and accuracies of the presented algorithms on a multiple alignment of twelve orthologous mammalian genomic sequences of $\sim$1 Mb each coming from the CFTR benchmark dataset (ENCODE Project Consortium, 2004).

2. THE INDEL MAXIMUM LIKELIHOOD PROBLEM

In this section we will give a precise definition for the Indel Maximum Likelihood Problem (IMLP). Consider a rooted binary phylogenetic tree $T = (V_T, E_T)$ with branch lengths $\lambda : V_T \to \mathbb{R}^+$. If n is the number of leaves of T, there are $n-1$ internal nodes and $2n-2$ edges.

Consider a multiple alignment A of n orthologous sequences corresponding to the leaves of the tree T. Since the only evolutionary events of interest here are insertions and deletions, A can be transformed into a binary matrix, where gaps are replaced by 0's and nucleotides by 1's. Let A_x be the row of the binarized alignment corresponding to the sequence at leaf x of T, and let $A_x[i]$ be the binary character at the i-th position of A_x. Assuming that the alignment A contains L columns, we add for convenience two extra columns, $A[0]$ and $A[L+1]$, consisting exclusively of 1's.

Definition 1 (Ancestral reconstruction). *Given a multiple alignment A of n extant sequences assigned to the leaves of a tree T, an* ancestral reconstruction *A^* is an extension of A that assigns a sequence $A^*_u \in \{0,1\}^{L+2}$ to each node u of T, and where $A^*_u = A_u$ whenever u is a leaf.*

The following restriction on the set of possible ancestral reconstructions is necessary in some contexts.

Definition 2 (Phylogenetically correct ancestral reconstruction). *An ancestral reconstruction A^* is* phylogenetically correct *if, for any $u, v, w \in V_T$ such that w is located on the path between u and v in T, we have $(A^*_u[i] = A^*_v[i] = 1) \implies (A^*_w[i] = 1)$.*

Requiring an ancestral reconstruction to be phylogenetically correct corresponds to assuming that any two nucleotides that are aligned in A have to be derived from a common ancestor, and thus that all the ancestral nodes between them have to have been a nucleotide. This prohibits aligned nucleotides to be the result of two independent insertions. Assuming that this property holds perfectly for a given alignment A is somewhat unrealistic, but, for mammalian sequences, good alignment heuristics have been developed—e.g., TBA (Blanchette et al., 2004b), MAVID (Bray and Pachter, 2004), and MLAGAN (Brudno et al., 2003)—and have been shown to be quite accurate (Blanchette et al., 2004b). In the future, we plan to relax this assumption, but, for now, we will concentrate only on finding phylogenetically correct ancestral reconstructions.

Since we are considering insertions and deletions affecting several consecutive characters, we delimit each operation by the positions s and e in the aligned sequences where it starts and ends. Let x and y be two nodes of the tree, where x is the parent of y. The pairwise alignment consisting of rows A^*_x and A^*_y is divided into a set of regions defined as follows (Fig. 2).

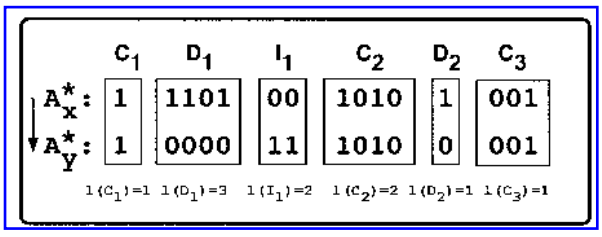


FIG. 2. Example of the partition of a pairwise alignment of A^*_x and A^*_y (where x is the parent of y) into deletions, insertions, and conservations. The length of each operation is given below it.

Definition 3 (Deletions, Insertions, Conservations, and Length). *Consider the pairwise alignment of A_x^* and A_y^*, and let $0 \leq s \leq e \leq L+1$.*

- *The region (s,e) is a* deletion *if (a) for all $i \in \{s,\ldots,e\}$, $A_y^*[i] = 0$, (b) $A_x^*[s] = A_x^*[e] = 1$, and (c) no region $(s',e') \supset (s,e)$ is a deletion (i.e., we only consider regions that are maximal).*
- *The region (s,e) is an* insertion *if (a) for all $i \in \{s,\ldots,e\}$, $A_x^*[i] = 0$, (b) $A_y^*[s] = A_y^*[e] = 1$, and (c) no region $(s',e') \supset (s,e)$ is an insertion.*
- *The region (s,e) is a conservation if (a) for all $i \in \{s,\ldots,e\}$, $A_x^*[i] = A_y^*[i]$ and (b) no region $(s',e') \supset (s,e)$ is a conservation.*
- *The* length *of region (s,e) is the number of non-trivial positions it contains: $l(s,e) = |\{s \leq i \leq e | A_x^*[i] \neq 0$ or $A_y^*[i] \neq 0\}|$.*

A pair of binary alignment rows A_x^* and A_y^* can thus be partitioned into a set of non-overlapping insertions, deletions, and conservations.

Definition 4 (Indel scenario). *The indel scenario defined by an ancestral reconstruction A^* is the set of insertions and deletions that occurred between the ancestral reconstructions at adjacent nodes in T.*

All that remains is to define an optimization criterion on A^*. Two main choices are possible: a parsimony criterion or a likelihood criterion.

2.1. The indel parsimony problem

The parsimony approach for the indel reconstruction problem has been introduced by Fredslund et al. (2004) and Blanchette et al. (2004a). In its simplest version, it attempts to find the phylogenetically correct ancestral reconstruction A^* that minimizes the total number of insertions and deletions defined by A^*:

$$\mathit{indelParsimony}(A^*) = \sum_{u,v:(u,v)\in E_T} |\{(s,e) : (s,e) \text{ is a deletion or an insertion from } A_u^* \text{ to } A_v^*\}|$$

The Indel Parsimony Problem is NP-Hard (Chindelevitch et al., 2006). Most authors have studied a weighted version of the *IPP* where the cost of indels depends linearly on their length (affine gap penalty). Blanchette et al. (2004a) proposed a greedy algorithm, and good exact heuristics have been developed (Chindelevitch et al., 2006; Fredslund et al., 2004). The limitation of these approaches is that they only give a single solution as output, and provide no measure of uncertainty of the various parts of the reconstruction. In contrast, a likelihood-based approach has the potential of providing a more accurate solution and a richer description of the set of possible solutions.

2.2. Indel maximum likelihood problem

In this section, we define the indel reconstruction problem in a probabilistic framework similar to the Thorne-Kishino-Felsenstein model (Thorne et al., 1992). To this end, we need to define the probability of transition between an alignment row A_x^* and its descendant row A_y^*. This probability will be defined as a function of the probability of the insertions, deletions, and conservations that happened from A_x^* to A_y^*.

Let $P_{DelStart}(\lambda(b))$ be the probability that a deletion starts at a given position in the sequence, along a branch b of length $\lambda(b)$, and let $P_{InsStart}(\lambda(b))$ be defined similarly for an insertion. We assume that these probabilities only depend on the length $\lambda(b)$ of the branch b along which they occur, but not on the position where the indel occurs. A reasonable choice is $P_{DelStart}(\lambda(b)) = 1 - e^{-\psi_D \lambda(b)}$ and $P_{InsStart}(\lambda(b)) = 1 - e^{-\psi_I \lambda(b)}$, for some deletion and insertion rate parameters ψ_D and ψ_I, but our algorithm allows for any other choice of these probabilities. Thus, the probability that none of the two events happens at a given position, which we call the probability of a conservation, is given by $P_{Cons}(\lambda(b)) = e^{-(\psi_D+\psi_I)\lambda(b)}$. We make the standard simplifying assumption that the length of a deletion follows a geometric distribution, where the probability of a deletion of length k is $\alpha_D^{k-1}(1-\alpha_D)$ and the probability of an insertion of length k is $\alpha_I^{k-1}(1-\alpha_I)$. One can thus see α_D (resp. α_I) as the probability of

extending a deletion (resp. insertion). This assumption, necessary to design a fast algorithm, holds relatively well for short indels, but fails for longer ones (Kent et al., 2003). Our algorithm allows the parameters α_D and α_I to depend on the branch b, but the results reported in Section 5 correspond to the case where α_D and α_I were held constant across the tree. The probability that alignment row A_x^* was transformed into alignment row A_y^* along branch b can be defined as follows:

$$\Pr(A_y^*|A_x^*, b) = \prod_{(s,e)\text{: deletion from } A_x^* \text{ to } A_y^*} P_{DelStart}(\lambda(b)) \cdot (\alpha_D^{l(s,e)-1}(1-\alpha_D))$$

$$\times \prod_{(s,e)\text{: insertion from } A_x^* \text{ to } A_y^*} P_{InsStart}(\lambda(b)) \cdot (\alpha_I^{l(s,e)-1}(1-\alpha_I))$$

$$\times \prod_{(s,e)\text{: conservation from } A_x^* \text{ to } A_y^*} (P_{Cons}(\lambda(b)))^{l(s,e)}$$

This allows us to formulate precisely the problem addressed in this paper:

INDEL MAXIMUM LIKELIHOOD PROBLEM

Given: A multiple sequence alignment A of n orthologous sequences related by a phylogenetic tree T with branch lengths λ, a probability model for insertions and deletions specifying the values of ψ_D, ψ_I, α_D, and α_I.

Find: A maximum likelihood phylogenetically correct ancestral reconstruction A^* for A, where the likelihood of A^* is:

$$L(A^*) = \prod_{b=(x,y)\in E_T} \Pr(A_y^*|A_x^*, b)$$

3. A TREE-HIDDEN MARKOV MODEL

In this section, we describe the tree hidden Markov model that is used to solve the IMLP. A tree-hidden Markov model (tree-HMM) is a probabilistic model that allows two processes to occur, one in time (related to the sequence history in a given column of A), and one in space (related to the changes toward the neighboring columns). Tree HMMs were introduced by Felsenstein and Churchill (1996) and Yang (1996) to improve the phylogenetic models that allows for variation among sites in the rate of substitution, and have since then been used for several other purpose—e.g., detecting conserved regions (Siepel et al., 2005) and predicting genes (Siepel and Haussler, 2004). Just as any standard HMM (Durbin et al., 1998), a tree-HMM is defined by three components: the set of states, the set of emission probabilities, and the set of transition probabilities.

3.1. States

Intuitively, each state corresponds to a different single-column indel scenario (although additional complications are described below). Given a rooted binary tree $T = (V_T, E_T)$ with n leaves, each state corresponds to a different labeling of the edges E_T with one of three possible events: I (for insertion), D (for deletion), or C (for conservation). The set $\mathcal{S}$ of possible states of the HMM would then be $\mathcal{S} = \{I, D, C\}^{2n-2}$. However, this definition is not sufficient to model certain biological situations (Fig. 3). We will use the '*' symbol to indicate that, along a certain branch $b = (x, y)$, no event happened because there was a base neither at node x nor at node y. This will happen in two situations: when edge b is a descendant of edge b' that was labeled with D (i.e., the base was deleted higher up the tree), and when there exists an edge b' that is not between b and the root and that is labeled with I (i.e., an insertion happened elsewhere in the tree). The fact that these extraneous events can potentially interrupt ongoing events along branch b means that the HMM needs to have a way to remember what event was actually

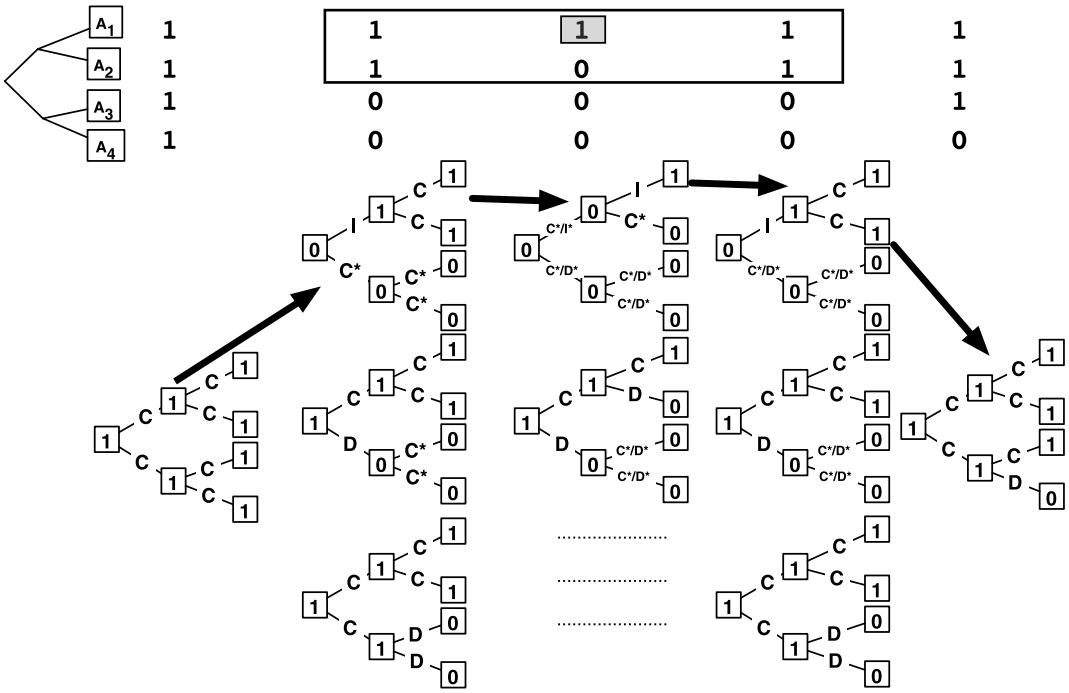


FIG. 3. The set of valid, non-zero probability states associated to the multiple alignment given at the top of the figure. When edges are labeled with more than one character (e.g., C^*, D^*), the tree represents several possible states. For the third column, not all possible states are shown. Arrows indicate one possible path through the tree-HMM. This path corresponds to two interleaved insertions, shown by two boxes in the alignment, illustrating the need for the I^* character.

going on along that branch. This transmission of memory from column to column is achieved by three special labels: I^*, D^*, and C^*, depending on whether the * regions is interrupting an insertion, deletion, or conservation. Thus, we have $\mathcal{S} \subseteq \{I, D, C, I^*, D^*, C^*\}^{2n-2}$. Although this state space appears prohibitively large (6^{2n-2}), the reality is that a number of these states cannot represent actual indel scenarios, and can thus be ignored. The following set of rules specify what states are valid.

Definition 5 (Valid states). *Given a tree $T = (V_T, E_T)$, a state s assigning a label $s(b) \in \{I, D, C, I^*, D^*, C^*\}$ to each branch $b \in E_T$ is* valid *if the two following conditions hold.*

- *(Phylogenetic correctness condition) There must be at most one branch b such that $s(b) = I$.*
- *(Star condition) Let $b \in E_T$, and let $anc(b) \subset E_T$ be the set of branches on the path from the root to b. Then $s(b) \in \{I^*, D^*, C^*\}$ if and only if $\exists b' \in anc(b)$ such that $s(b') = D$ or $\exists b' \in (E_T \setminus anc(b))$ such that $s(b') = I$.*

The number of valid states on a complete balanced phylogenetic tree with n leaves is $O(n \cdot 3^{2n})$ (the number is dominated by states that have an "I" on a branch leading to a leaf, which leaves all other $2n-3$ edges free to be labeled with either C^*, D^*, or I^*). Although this number remains exponential, it is significantly better than the 6^{2n-2} valid and invalid states.

3.2. Emission probabilities

In an HMM, each state emits one symbol, according a certain emission probability distribution. In our tree-HMMs, each state emits a collection of symbols, corresponding to the set of characters obtained at the leaves of T when indel scenario s occurs. Intuitively, we can think of a state as emitting an alignment column. The following definition formalizes this.

TABLE 1. EDGE TRANSITION TABLE $\rho(s'(e)|s(e), b)$

$s(e) \setminus s'(e)$	C	D	I	C^*	D^*	I^*
C	$P_{Cons}(\lambda(b))$	$P_{DelStart}(\lambda(b))$	$P_{InsStart}(\lambda(b))$	1	0	0
D	$(1-\alpha_D)P_{Cons}(\lambda(b))$	α_D	$(1-\alpha_D)P_{InsStart}(\lambda(b))$	0	1	0
I	$(1-\alpha_I)P_{Cons}(\lambda(b))$	$(1-\alpha_I)P_{DelStart}(\lambda(b))$	α_I	0	0	1
C^*	$P_{Cons}(\lambda(b))$	$P_{DelStart}(\lambda(b))$	$P_{InsStart}(\lambda(b))$	1	0	0
D^*	$(1-\alpha_D)P_{Cons}(\lambda(b))$	α_D	$(1-\alpha_D)P_{InsStart}(\lambda(b))$	0	1	0
I^*	$(1-\alpha_I)P_{Cons}(\lambda(b))$	$(1-\alpha_I)P_{DelStart}(\lambda(b))$	α_I	0	0	1

Notice that ρ is not a transition probability matrix, since its rows sum to more than one.

Definition 6. *Let s be a valid state for tree $T = (V_T, E_T)$ with root r. Then, we define the output of state s as a function $O_s : V_T \to \{0, 1\}$ with the following recursive properties:*

1. $O_s(root) = \begin{cases} 0, & \text{if } \exists x \in V_T \text{ such that } s(x) = I \\ 1, & \text{otherwise} \end{cases}$.

2. *Let $e = (x, y) \in E_T$, with x being the parent of y. Then,*

$$O_s(y) = \begin{cases} 0, & \text{if } s(e) = D \\ 1, & \text{if } s(e) = I \\ O_s(x), & \text{otherwise} \end{cases}$$

Let C be an alignment column (i.e., an assignment of 0 or 1 to each leaf in T). We then have the following degenerate emission probability for state s:

$$\Pr_e(C|s) = \begin{cases} 1 \text{ if } O_s(x) = C(x) \text{ for all } x \in leaves(T) \\ 0 \text{ otherwise} \end{cases}$$

Thus, each state s can emit a single alignment column C. However, many different states can emit the same column.

Missing data. In presence of missing characters among the input sequences, the emission probability can be adapted such that the equality between $O_s(x)$ and $C(x)$ is assessed according to 0's and 1's in $C(x)$ only. It is worth noting that missing characters are different to gaps noted by $-$. Hence, the presence of missing data increases the number of states for a given column.

3.3. Transition probabilities

The last component to be defined is the set of transition probabilities of the tree-HMM. The probability of transition from state s to state s', $\Pr_t(s'|s)$, is a function of the set of events that occurred along the edges of T. Intuitively, $\Pr_t(s'|s)$ describes the probability of the single-column indel scenario s', given that scenario s occurred at the previous column. This transition probability is a function of insertions and deletions that started between the two columns, of those that were extended going from one column to the next. Specifically, we have $\Pr_t(s'|s) = \prod_{b \in E_T} \rho(s'(e)|s(e), b)$, where ρ is given in Table 1.

4. TREE-HMM PATHS, ANCESTRAL RECONSTRUCTION, AND ASSESSING UNCERTAINTY

We now show how the tree-HMM described above allows us to solve the IMLP. Consider a multiple alignment A of length L on a tree T. A path π in the tree-HMM is a sequence of states $\pi = \pi_0, \pi_1, ..., \pi_L, \pi_{L+1}$.

Based on standard HMM theory, we get:

$$\Pr(\pi, A) = \Pr(\pi_0, A_0) \prod_{i=1}^{L+1} \Pr_e(A[i]|\pi_i) \cdot \Pr_t(\pi_i|\pi_{i-1})$$

Figure 3 gives an example of an alignment with some of the non-zero probability paths associated.

Theorem 1. *Consider an alignment A on tree T. Then $\pi^* = argmax_\pi \Pr(\pi, A)$ yields the most likely indel scenario for A, and a maximum likelihood ancestral reconstruction A^* is obtained by setting $A^*_u[i] = O_{\pi^*_i}(u)$.*

Proof. It is simple to show that for any ancestral reconstruction $\hat{A}$ for A, we have $L(\hat{A}) = \Pr(\pi, A)$, where π is the path corresponding to $\hat{A}$. Thus, maximizing $\Pr(\pi, A)$ maximizes $L(\hat{A})$. ■

4.1. Computing the most likely path

To compute the most likely path π^* through a tree-HMM, we adapted the standard Viterbi dynamic programming algorithm (Viterbi, 1967). Let $X(i, k)$ be the joint likelihood of the most probable path ending at state k for the i first columns of the alignment. Let $c \in \mathcal{S}$ be the state made of C's on all edges of T. Since the dummy column $A[0]$ consists exclusively of 1's, c is the only possible initial state. For any i between 0 and $L + 1$ and for any valid state $s \in \mathcal{S}$, we can compute $X(i, s)$ as follows:

$$X(i,s) = \begin{cases} 1, & \text{if } i = 0 \text{ and } s = c \\ 0, & \text{if } i = 0 \text{ and } s \neq c \\ \Pr_e(A[i]|s) \cdot \max_{s' \in \mathcal{S}}(X(i-1, s') \cdot \Pr_t(s|s')), & \text{if } i > 0 \end{cases}$$

Finally, π^* is obtained by tracing back the dynamic programming, starting from entry $X(L + 1, c)$. To ensure numerical stability, we use a log transformation and scaling of probabilities as described by Durbin et al. (1998).

The running time of a naive implementation of the Viterbi algorithm is $O(|\mathcal{S}|^2 L)$, which quickly becomes impractical as the size of the tree T grows. However, we can make this computation practical for moderately large trees and for long sequences. Even though the number of states is exponential in the number of sequences, most alignment columns can only be generated with non-zero probability by a much more manageable number of states. Given an alignment A, it is possible to compute, for each column $A[i]$, the set S_i of valid states that can emit $A[i]$ with non-zero probability. For instance, an alignment column with only 1's will lead to only one possible state, independently of the number taxa of n. The set S_i can be constructed using a bottom approach presented in Algorithm 1. More states can be discarded by using the fact that the transition probability between most pairs of states is zero. We can thus remove from S_i any state s that is such that the transition to s from any state in S_{i-1} has probability zero. Proceeding from left to right, we get $S'_0 = S_0$, and $S'_i = \{s \in S_i | \exists t \in S'_{i-1} \text{ s.t. } \Pr_t(s|t) > 0\}$, where $S'_i \subseteq S_i$. For instance, if, in all states of S_{i-1}, an edge e is labeled by deletion D, then none of the states in S_i can have edge e labeled with C^* or I^*. This yields a large improvement for alignment regions consisting of a number of adjacent positions with a base in only one of the n species and ensures that the algorithm will be practical for relatively large number of sequences (see Section 5).

4.2. Assessing uncertainties of the ancestral reconstruction

A significant advantage of the likelihood approach over the parsimony approach is that it allows evaluating the uncertainty related to certain aspects of the reconstruction. For example, it is useful to be able to compute the probability that a base was present at a given position i of a given ancestral node u: $\Pr(A^*_u[i] = 1|A) = \sum_{s \in \mathcal{S}: O_s(u)=1} \Pr(\pi_i = s|A)$. This allows the computation of the probability of making an incorrect prediction at a given position of a given ancestor. The forward-backward is a standard HMM algorithm to compute $\Pr(\pi_i = s|A)$ (Durbin et al., 1998). The optimizations developed for the Viterbi algorithm can be trivially adapted to the Forward-Backward algorithm.

Algorithm 1 buildValidState(node *root*, C)

Require: *root*: a tree node, C: an alignment column.
Ensure: Set of valid, non-zero probability states for C.

```
1: if root is a leaf then
2:    return list of possible operations according to the character at that leaf
3: else
4:    leftList = buildValidState(root.left, C)
5:    rightList = buildValidState(root.right, C)
6:    return mergeSubtrees(leftList, rightList, root)
7: end if
```

Algorithm 2 mergeSubtrees(StateList *leftList*, StateList *rightList*, node *root*)

Require: *leftList* and *rightList*: the lists of partial states, *root*: a tree node.
Ensure: Set of valid, non-zero probability states combining elements in *leftList* and *rightList*.

```
1: mergedList ← emptyList
2: for all partial states l in leftList do
3:    for all partial states r in rightList do
4:       if compatible(l, r) == true then
5:          m = merge(l, r)
6:          if root == initialroot then
7:             mergedList.add(m)
8:          else
9:             for op ∈ {C, D, I, C*, D*, I*} do
10:               if isPossibleUpstream(m,op) then
11:                  mergedList.add(addAncestorBranch(m,op))
12:               end if
13:            end for
14:         end if
15:      end if
16:   end for
17: end for
18: return mergedList
```

5. RESULTS OF THE EXACT METHOD

Our tree-HMM algorithm was implemented as a C program that is available upon request. The program was applied to a ~700-kb region of the CFTR locus on chromosome 7 of human, together with orthologous regions in 11 other species of mammals: chimp, macaque, baboon, mouse, rat, rabbit, cow, dog, Rodrigues fruit bat (rfbat), armadillo, and elephant[1] (ENCODE Project Consortium, 2004). This locus is representative of the whole genome, and contains coding, intergenic regions, and intronic regions. The multiple alignment of these regions, computed using TBA (Blanchette et al., 2004b; Miller, 2006), contains 1,000,000 columns. To simplify the calculations, consecutive alignment columns with the same gap structure were assumed to have undergone the same evolutionary scenario and were thus merged into a single "meta-column" we called an alignment *region*. Our alignment consisted of 123,917 such regions. Thus, during the execution of the Viterbi or Forward-Backward algorithm, the states are computed for each region instead of for

[1]In the case of cow, armadillo, and elephant, the sequence is incomplete and a small fraction of the bases are missing.

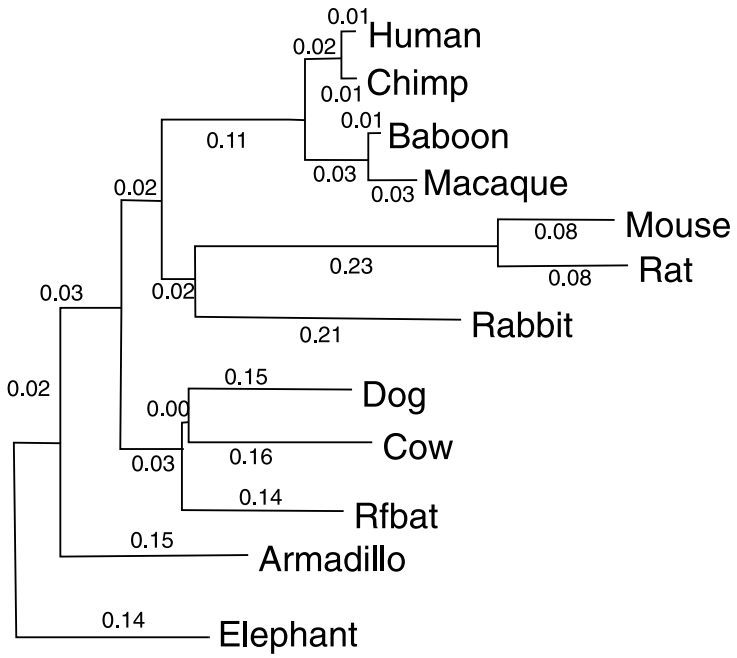


FIG. 4. Phylogenetic tree for the twelve species studied in this paper.

each individual column, adapting the transition probabilities as a function of the width of each region. The phylogenetic tree used for the alignment and for the reconstruction is shown in Figure 4. The branch lengths are based on substitution rates estimated on a genome-wide basis (Miller, 2006). For illustrative purposes, and similarly to the empirical values obtained by Kent et al. (2003), the parameters of the indel model were set as follows: $\psi_D = 0.05, \psi_I = 0.05, \alpha_D = 0.9$, and $\alpha_I = 0.9$. However, we find that the ancestral reconstructions and confidence levels are quite robust with respect to these parameters (data not shown).

We first compared the maximum likelihood ancestral reconstruction found using our Viterbi algorithm to the ancestors inferred using the greedy algorithm of Blanchette et al. (2004a). Table 2 shows the degree

TABLE 2. PERCENTAGE OF ALIGNMENT COLUMNS WHERE THERE IS AGREEMENT BETWEEN THE ANCESTOR RECONSTRUCTED BY THE GREEDY ALGORITHM OF BLANCHETTE ET AL. (2004A) AND THAT PREDICTED BY OUR MAXIMUM-LIKELIHOOD ALGORITHM

Ancestor	*Percentage of agreement*
Mou+Rat	99.8181
Hum+Chi	99.9467
Bab+Mac	99.7275
Mou+Rat+Rab	99.8181
Hum+Chi+ Bab+Mac	99.7157
Hum+Chi+Bab+Mac+Mou+Rat+Rab	99.3901
Cow+Dog	99.917
Cow+Dog+Bat	99.8218
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat	99.0511
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat+Arm	93.6531
Hum+Chi+Bab+Mac+Mou+Rat+Rab+Cow+Dog+Bat+Arm+Ele	84.9413

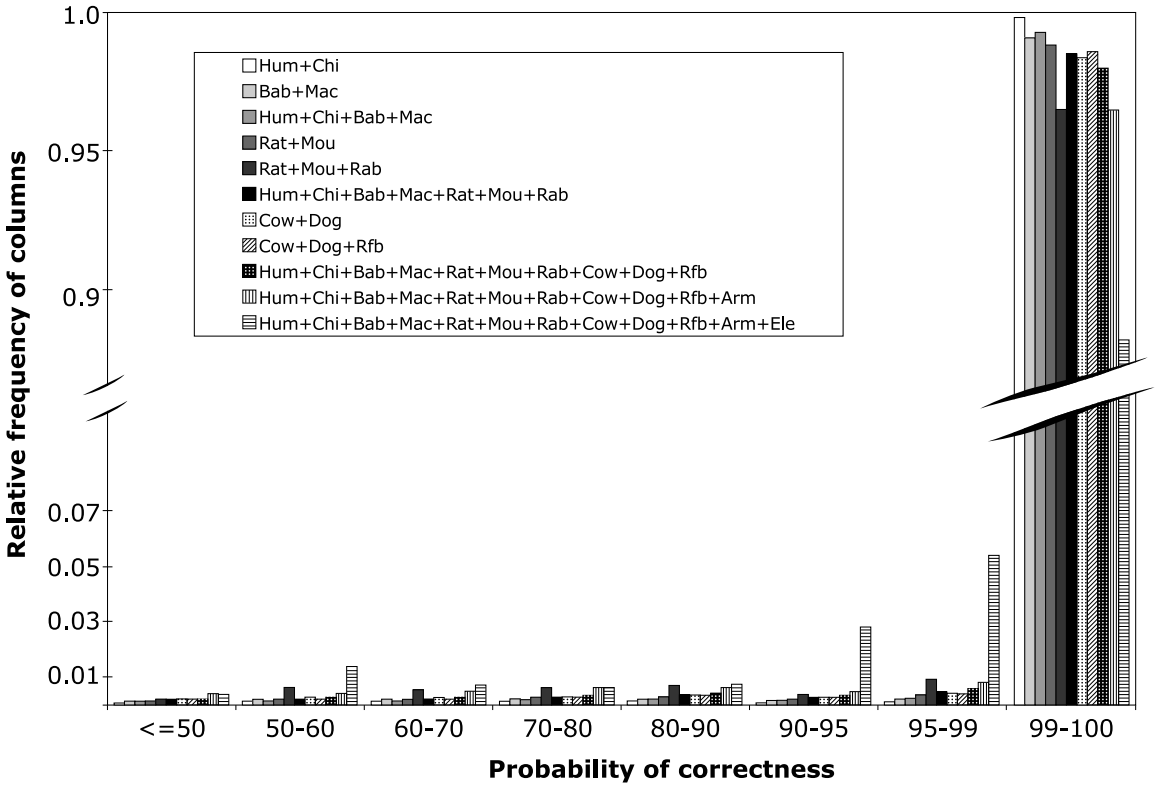


FIG. 5. Distribution of the confidence levels, over all 123,917 alignment regions, for each ancestor. The vast majority of the ancestral positions are reconstructed with a probability of correctness above 99% (assuming the correctness of the alignment).

of agreement between the two reconstructed ancestors, for each ancestral node. We observe that both methods agree to a very large degree, with most ancestors yielding more than 99% agreement. The most disagreement concerns the ancestor at the root of the eutherian tree, which, in the absence of an outgroup, cannot be reliably predicted by any method. We expect that in most other cases of disagreement, the maximum likelihood reconstruction is the most likely to be correct, although the opposite may be true in case of gross model violations (Hudek and Brown, 2005).

The main strength of the likelihood-based method is its ability to measure uncertainty, using the forward-backward algorithm, something that no previous method allowed. Assuming a phylogenetically correct alignment and a correct indel model, the probability that the maximum posterior probability reconstruction is correct is simply given by $\max\{\Pr(A_u^*[i] = 1|A), 1 - \Pr(A_u^*[i] = 1|A)\}$. For example, if $\Pr(A_u^*[i] = 1|A) = 0.3$, then the maximum posterior probability reconstruction would predict $A_u^*[i] = 0$, and would be right with probability 0.7. Figure 5 shows the distribution of this probability of correctness, for each ancestral node in the tree, over all regions of the alignment. We observe, for example, that 98% of the positions in the Boreoeutherian ancestor (the human+chimp+baboon+macaque+mouse+rat, cow+dog+rfbat ancestor, living approximately 75 million years ago), are reconstructed with a confidence level above 99%.[2] The ancestor that is the easiest to reconstruct confidently is obviously the human-chimp ancestor, where less than 0.14% of the regions have a confidence level below 99%. Again, the root of the tree is the node that is the most difficult to reconstruct confidently. Overall, this shows that most positions

[2]We need to keep in mind, though, that these numbers assume the correctness of the multiple alignment, as well as that of the branch lengths and indel probability model, so that they do not reflect the true correctness of the reconstructed ancestor.

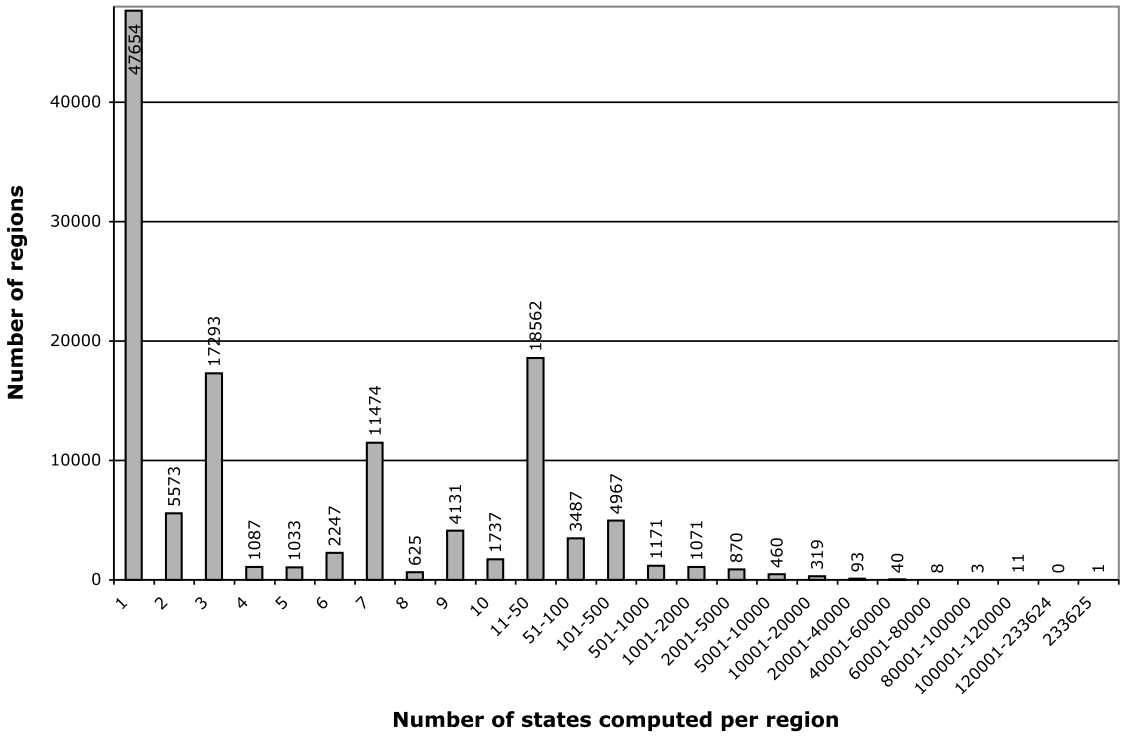


FIG. 6. Distribution of the number of states considered ($|S'_i|$), over all 123,917 regions.

of most ancestral nodes can be reconstructed very accurately, and that we can identify the few positions where the reconstruction is uncertain.

A potential drawback of the tree-HMM method is that its running time is, in the worst case, exponential in the number of sequences being compared. However, the optimizations described in this paper greatly reduce the number of states that need to be considered at each position, so the algorithm remains quite fast. Our optimized Viterbi algorithm produced its maximum likelihood ancestral predictions on the 12-species, 1,000,000 column alignment in 7 hours on a Powerbook G5 machine, while the forward-backward algorithm produced an output after approximately double of that time. Figure 6 shows the distribution of the number of states that were actually considered, per alignment column. Most alignment columns are actually associated to less than 100 states. However, a small number of columns are associated to a very large number of states (15 regions have more than 100,000 states). Fortunately, these columns are rarely consecutive, so the incurred running time is not catastrophic for small number of species. However, to be applicable to complete genomes and to scale up to the more than 20 mammalian genomes that will soon be available, our algorithm requires further optimizations. These optimizations move away from an exact algorithm, toward approximation algorithms.

6. HEURISTIC ALGORITHM FOR THE IMLP

For each region i of the alignment and each possible state $s \in S'_i$, the exhaustive method considers all possible states for the next column, even though the Viterbi value $X(i,s)$ of some current state s may be far away from the maximal Viterbi value at that position, $\max_{s' \in S'_i} X(i,s')$. These states are less likely to be eventually chosen in the best path of the tree-HMM. Hence, to reduce the number of states created and reduce computation time, only states near the maximum Viterbi value are used to compute states for the next column. Thus, for region i, we distinguish between created states S'_i and used states $R_i \subseteq S'_i$, where only the second set will be involved in the creation of the states of the next column and in their Viterbi calculation. For position i, state $s \in S'_i$ is retained in R_i if and only if $\log_2(\frac{\max_{s'} X(i,s')}{X(i,s)}) < t$, for some

fixed threshold t. We note that this is equivalent to setting $X(i, s)$ to zero for each $s \in S - R$. A similar heuristic can easily be applied to the Forward-Backward algorithm. If t is sufficiently large, the loss in accuracy should be minimal for both algorithms, as will be shown next.

We computed the indels scenarios of the data sets presented in Section 5 by using different values for the threshold t. The approximate Viterbi algorithm was run using $t = 0, 1, 3, 5, 7, 9, 10, 20, 100$, and $+\infty$. Note that setting $t = 0$ results in a "greedy" algorithm that only considers the maximum Viterbi value at each position, while $t = +\infty$ give the original, optimal Viterbi algorithm. Figure 7 shows the number of states created (average of $|S'_i|$) and used (average of $|R_i|$) for all values of t, as well as the resulting running time. For small values of t, e.g., $t \leq 3$, only a handful of states are used, resulting in a very fast execution (less than 3 minutes). The average of number of states created increases relatively quickly with t, while the number of states used remains quite low (44.34 for $t = 100$). The average number of states created for $t = 20$ is about the same as the average number of states of the exact algorithm (see Figure 7), which shows that the used states are sufficient to give the necessary information to generate most valid states for next columns.

Even though the average number of states created and used for $0 \leq t \leq 5$ is very low, the indels scenarios produced are very similar to the best scenario obtained by the exact method (see Table 3). We note that, for $t = 5$, the agreement with the exact algorithm is more than 99.99% for all the ancestors, while the running time is reduced by a factor of ten, and by a factor of one hundred for $t = 3$. For $t \geq 9$, the heuristic gives the optimal scenario, while still yielding a 5-fold speed-up. All values of t tested gave solutions that agreed with the optimal solution better than the solution produced by the greedy algorithm of Blanchette et al. (2004a). Finally, we note that, while our optimal Viterbi and Forward-Backward algorithms are limited to 12 to 15 species, our heuristic allows the inference of near-optimal solutions for much larger alignments. When run on a 1,000,000 column alignment of 28 species of vertebrates, our heuristic with $t = 3$ produced a solution in less than two hours. Since the exact algorithm cannot be run on such a large data set, it is

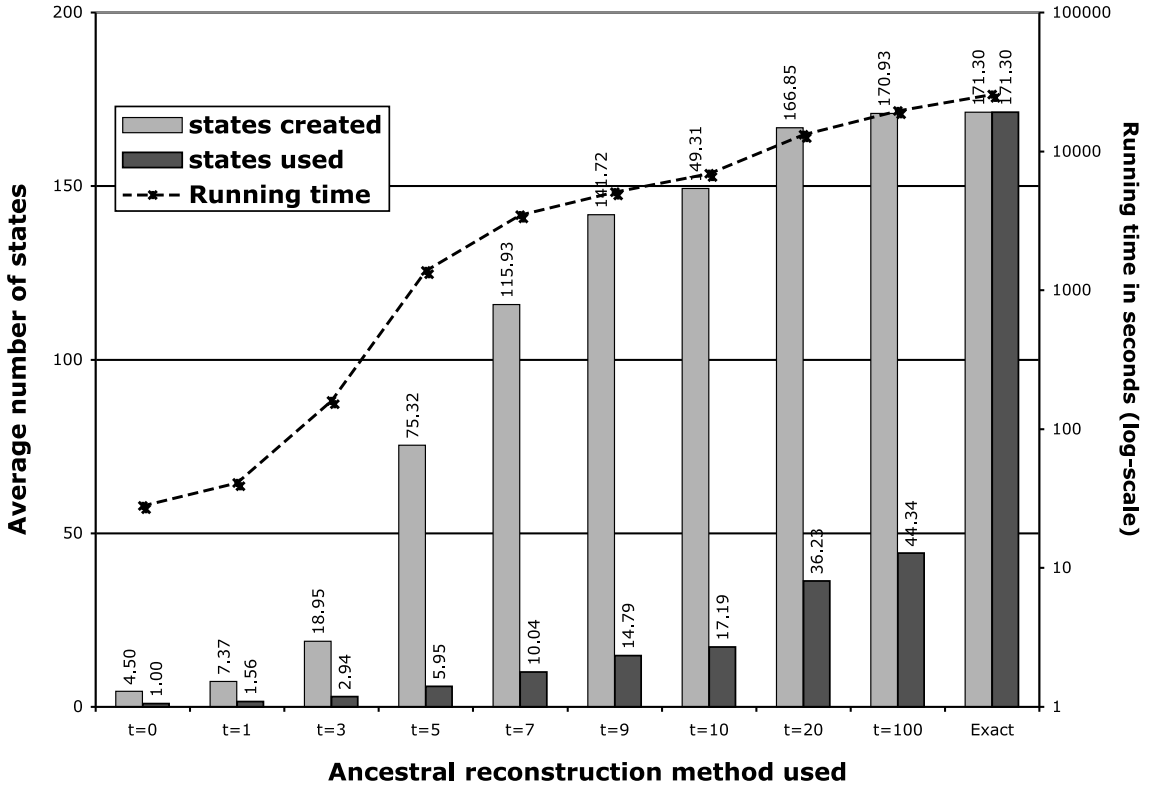


FIG. 7. Average, over all alignment regions, of the number of states created (S'_i) and used (R_i), for the different values of the cutoff t. Running times (in seconds) are plotted with the log-scale shown on the right.

TABLE 3. PERCENTAGE OF ALIGNMENT COLUMNS WHERE THERE IS DISAGREEMENT BETWEEN THE ANCESTOR RECONSTRUCTED BY THE EXACT MAXIMUM-LIKELIHOOD ALGORITHM AND THE HEURISTIC WITH DIFFERENT VALUES FOR THE CUTOFF t

Ancestor	$t = 0$	$t = 1$	$t = 3$	$t = 5$	$t = 7$	$t > 9$
Mou+Rat	0.030	0.012	0.003	0.002	0.001	0
Hum+Chi	0.020	0.004	0.001	0.001	0.001	0
Bab+Mac	0.003	0.003	0.002	0.002	0.002	0
Mou+Rat+Rab	0.160	0.073	0.008	0.003	0.002	0
Hum+Chi+ Bab+Mac	0.060	0.041	0.011	0.002	0.002	0
Hum+Chi+Bab+Mac+Mou+Rat+Rab	0.160	0.070	0.018	0.006	0.004	0
Cow+Dog	0.070	0.032	0.006	0.002	0.001	0
Cow+Dog+Bat	0.080	0.049	0.013	0.002	0.001	0
Hum+Chi+Bab+Mac+Mou+Rat+ Rab+Cow+Dog+Bat	0.170	0.095	0.017	0.005	0.004	0
Hum+Chi+Bab+Mac+Mou+Rat+ Rab+Cow+Dog+Bat +Arm	0.100	0.048	0.010	0.003	0.002	0
Hum+Chi+Bab+Mac+Mou+Rat+ Rab+Cow+Dog+Bat +Arm+Ele	0.010	0.004	0	0	0	0

We emphasize that the numbers quoted are percentages, so, for example, with $t = 0$, the Mouse+Rat ancestor agrees with the optimal solution at 99.97% of the alignment columns.

difficult to estimate the quality of the solution obtained but, based on our experience on the smaller data set (Table 3), we expect a very high accuracy even at such a stringent cutoff.

7. DISCUSSION

The method developed here allows predicting maximum likelihood indel scenarios and their resulting ancestral sequences for large alignments. Furthermore, it allows the estimation of the probability of error in any part of the prediction, using the forward-backward algorithm. Integrated into the pipeline for whole-genome ancestral reconstruction, it will improve the quality of the predictions and allow richer analyses. The main weakness of our approach is that it assumes that a phylogenetically correct alignment and an accurate phylogenetic tree are given as input. While many existing multiple alignment programs have been shown to be quite accurate on mammalian genomic sequences (including non-functional or repetitive regions) (Blanchette et al., 2004b), it has also been shown that a sizeable fraction of reconstruction errors is due to incorrect alignments (Blanchette et al., 2004a). Ideally, one would include the optimization of the alignment directly in the indel reconstruction problem, as originally suggested by Hein (1989). However, with the exception of statistical alignment approaches (Lunter et al., 2003), which remain too slow to be applicable on a genome-wide scale, genomic multiple alignment methods do not treat indels in a probabilistic framework. We are thus investigating the possibility of using the method proposed here to detect certain types of small-scale alignment errors, and to suggest corrections.

When predicting ancestral genomic sequences, it is very important to be able to quantify the uncertainty with respect to certain aspects of the reconstruction. Our forward-backward algorithm calculates this probability of error for each position of each ancestral species. However, errors in adjacent columns are not independent: if position i is incorrectly reconstructed, it is very likely that position $i + 1$ will be wrong too. We are currently working on models to represent this type of correlated uncertainties. This new type of representation will play an important role in the analysis and visualization of ancestral reconstructions.

Finally, it will be important to assess the results given by the heuristic so that the cutoff value t is chosen appropriately for the data at hand. For example, the heuristic could be applied iteratively by increasing the cutoff until a stationary likelihood score is reached. This heuristic will be useful to reconstruct the indel scenarios for data sets containing more than 20 taxa and could be easily applied to the large number of mammalian genomes that are about to be completely sequenced.

ACKNOWLEDGMENTS

We thank Éric Gaul, Eric Blais, Adam Siepel, and the group of participants to the First Barbados Workshop on Paleogenomics for their useful comments. We thank Webb Miller and David Haussler for providing us with the sequence alignment data. A.B.D. is an NSERC fellow.

REFERENCES

Benner, S. 2002. The past as the key to the present: resurrection of ancient proteins from eosinophils. *Proc. Natl. Acad. Sci. USA* 99, 4760–4761.

Blanchette, M., Green, E.D., Miller, W., et al. 2004a. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.* 14, 2412–2423.

Blanchette, M., Kent, W.J., Riemer, C., et al. 2004b. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* 14, 708–715.

Bray, N., and Pachter, L. 2004. MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.* 14, 693–699.

Brudno, M., Do, C.B., Cooper, G.M., et al. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* 13, 721–731.

Chindelevitch, L., Li, Z., Blais, E., et al. 2006. On the inference of parsimonious indel evolutionary scenarios. *J. Bioinform. Comput. Biol.* 4, 721–744.

Durbin, R., Eddy, S., Krogh, A., et al. 1998. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia of DNA elements) project. *Science* 306, 636–640.

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

Felsenstein, J., and Churchill, G. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* 13, 93–104.

Fitch, W.M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool.* 20, 406–416.

Fredslund, J., Hein, J., and Scharling, T. 2004. A large version of the small parsimony problem. *Proc. 4th WABI*, 417–432.

Gaucher, E., Thomson, M., Burgan, M., et al. 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature* 425, 285–288.

Hein, J. 1989. A method that simultaneously aligns, finds the phylogeny and reconstructs ancestral sequences for any number of ancestral sequences. *Mol. Biol. Evol.* 6, 649–668.

Hudek, A., and Brown, D. 2005. Ancestral sequence alignment under optimal conditions. *BMC Bioinform.* 6, 1–14.

Kent, W.J., Baertsch, R., Hinrichs, A., et al. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* 100, 11484–11489.

Kim, J., and Sinha, S. 2007. Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. *Bioinformatics* 23, 289–297.

Krishnan, N., Seligman, H., Stewart, C., et al. 2004. Ancestral sequence reconstruction in primate mitochondrial DNA: Compositional bias and effect on functional inference. *Mol. Biol. Evol.* 21, 1871–1883.

Lunter, G., Miklos, I., Song, Y., et al. 2003. An efficient algorithm for statistical multiple alignment on arbitrary phylogenetic trees. *J. Comput. Biol.* 10, 869–889.

Miller, W. 2006. Personal communication.

Pauling, L., and Zuckerkandl, E. 1963. Molecular "restoration studies" of extinct forms of life. *Acta Chem. Scand.* 17, 9–16.

Rivas, E. 2005. Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinform.* 6, 63.

Siepel, A., Bejerano, G., Pedersen, J.S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050.

Siepel, A., and Haussler, D. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* 11, 413–428.

Thorne, J.L., Kishino, H., and Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* 33, 114–124.

Thorne, J., Kishino, H., and Felsenstein, J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.* 34, 3–16.

Viterbi, A. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory* 13, 260–269.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.* 11, 367–372.

Address reprint requests to:
Dr. Mathieu Blanchette
McGill Centre for Bioinformatics and School of Computer Science
McGill University
3775 University St.
Montréal, Québec, H3A 2B4, Canada

E-mail: blanchem@mcb.mcgill.ca

From: "Ballen, Karen" <KBallen@liebertpub.com>
Subject: **FW: Liebert Online feedback**
Date: April 29, 2009 2:56:42 PM GMT-04:00
To: <diallo.abdoulaye@uqam.ca>

Dear Diallo:
Copyright permission is granted for this request.
Kind regards,
Karen Ballen
Manager, Reprint Department

From: Abdoulaye Banire Diallo [mailto:diallo.abdoulaye@uqam.ca]
Posted At: Tuesday, April 28, 2009 11:26 AM
Posted To: Liebert Online
Conversation: Liebert Online feedback
Subject: Liebert Online feedback

System information:

User: not logged in
Institution(s):
Date/time: Tue Apr 28 08:25:57 PDT 2009
Previous page: http://www.liebertonline.com/doi/abs/10.1089/cmb.2007.A006
Browser/OS: Mozilla/5.0 (Macintosh; U; Intel Mac OS X 10.5; en-US; rv:1.9.0.8) Gecko/2009032608 Firefox/3.0.8
IP Address: 132.208.137.068

User entered information:

Name: Abdoulaye Banire Diallo
Institution/affiliation: =?UTF-8?Q?UniversitÃ©_du_QuÃ©bec_?= =?UTF-8?Q?Ã _MontrÃ©al Depa?= =?UTF-8?Q?rtment:_Informatique City/town?= =?UTF-8?Q?:_ Country:_Canada ZIP/postal_c?= =?UTF-8?Q?ode:_H3C_3P8 Customer_number:_?= =?UTF-8?Q? E-mail:_diallo.abdoulaye@uqam.ca Question_regarding:_Other J?= =?UTF-8?Q?ournal:_Journal_of_Computational_Biology Question: Dear_publishers,?I would like to include my published paper: "Abdoulaye Banire Diallo, Vladimir Makarenkov, Mathieu Blanchette. Journal of Computational Biology. May 2007, 14(4): 446-461. doi:10.1089/cmb.2007.A006." into my Ph.D. thesis manuscript. For this purpose, I need a permission from the editor. Thus, it will be very helpful to have an authorization from you. Thank you and Best regards Abdoulaye BanirÃ© Diallo

---

Send copy: yes

From: "Permissions Europe/NL" <Permissions.Dordrecht@springer.com>
Subject: **RE: a permission to include my published papers in my thesis**
Date: May 5, 2009 3:30:36 AM GMT-04:00
To: "Abdoulaye Banire Diallo" <banire@gmail.com>

'dear dr. Banier Diallo ,
With reference to your request (copy herewith) to reprint material on which Springer Science and Business Media controls the copyright, our permission is granted, free of charge, for the use indicated in your enquiry.
This permission

- *allows you non-exclusive reproduction rights throughout the World.*
- *permission includes use in an electronic form, provided that content is*
  ** password protected;*
  ** at intranet;*
- *excludes use in any other electronic form. Should you have a specific project in mind, please reapply for permission.*
- *requires a full credit (Springer/Kluwer Academic Publishers book/journal title, volume, year of publication, page, chapter/article title, name(s) of author(s), figure number(s), original copyright notice) to the publication in which the material was originally published, by adding: with kind permission of Springer Science and Business Media.*

** The material can only be used for the purpose of defending your dissertation, and with a maximum of 40 extra copies in paper.*
Permission free of charge on this occasion does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

Berendina van Straalen
Springer
Head of Rights and Permissions
Special Licensing Department
Van Godewijckstraat 30 | 3311 GX Dordrecht
P.O.Box 17 | 3300 AA Dordrecht
The Netherlands
Tel. +31 78 657 62 10
Fax +31 78 657 63 77
www.springer.com

2009 FAIRS:
20-22 April - London
14-16 October – Frankfurt
Book fairs - permissions.bookfairs@springer.com
Translation rights - translations.dordrecht@springer.com
offers: http://www.springer.com/rights?SGWID=0-122-0-0-0
Special Licensing - TPL.dordrecht@springer.com
Permissions – permissions.dordrecht@springer.com

From: Abdoulaye Banire Diallo [mailto:banire@gmail.com]
Sent: donderdag 30 april 2009 19:45
To: Permissions Europe/NL
Cc: Abdoulaye Banire Diallo
Subject: a permission to include my published papers in my thesis

Dear publishers,
I would like to include my following 3 published papers into my Ph.D. thesis manuscript. For this purpose, I need a permission from the editor. Thus, it will be very helpful to have these authorizations from you. Here are the lists of the corresponding articles:

1) Abdoulaye Banire Diallo, Vladimir Makarenkov, Mathieu Blanchette (2006): Finding the most likely indel scenarios. *G. Bourque and N. El-Mabrouk (Eds.): RECOMB-CG2006, LNBI* 4205, pp.171–185, 2006

2) Abdoulaye Banire Diallo, Vladimir Makarenkov, Mathieu Blanchette and

Francois-Joseph Lapointe (2006): A new efficient method for assessing missing nucleotides in DNA sequences in the Framework of a generic evolutionary model. *Proceedings of the meeting of the International Federation of Classification Societies 2006, Data Science and Classification. eds Batagelj, V., Bock, H.H., Ferligo j, A., Ziberna, A.*, Springer Verlag, Ljublijana, 333-340.

3) Blanchette, M., Diallo, A.B., Green, E. D., Miller, W. and Haussler, D. (2007): Computational reconstruction of ancestral DNA sequences. Chapter 11 of the book : Methods in Molecular Biology: Phylogenomics. *Edited by: W. J. Murphy, Humana Press Inc., Totowa*, NJ, 171-184.

Thank you for quick reply. I need to submit it before May 4th 2009.

My fax number is :1 514 987 8477 if necessary

I am living in Montreal, Qc, Canada

Abdoulaye Baniré Diallo

From: "Erin Buck" <emb@ams.org>
Subject: **RE: written authorization to include one of my published paper in my Ph.D. thesis**
Date: May 5, 2009 11:51:15 AM GMT-04:00
To: <banire@mcb.mcgill.ca>
Cc: <banire@gmail.com>

Dear Mr. Baniré Diallo,

I am writing from the AMS regarding your request to include your paper, "Algorithms for detecting complete and partial horizontal gene transfers: theory and practice." (*CRM Proceedings & Lecture Notes*, vol. 45, 2008, pp. 159-179), in your Ph.D. thesis.

Permission from the AMS is not needed in this case, provided that original publication by the AMS is appropriately credited. This is outlined in clause 4 of the Consent to Publish form which is completed by AMS authors prior to publication. Clause 4 of the Consent to Publish form states the following :

4. The Work may be reproduced by any means for educational and scientific purposes by the Author(s) or by others without fee or permission, with the exception that reproduction by services that collect fees for delivery of documents may be licensed only by the Publisher. The Author(s) may use part or all of this Work or its image in any future works of his/her (their) own. In any reproduction, the original publication by the Publisher must be credited in the following manner: "First published in [Publication] in [volume and number, or year], published by the American Mathematical Society," and the copyright notice in proper form must be placed on all copies. Any publication or other form of reproduction not meeting these requirements will be deemed to be unauthorized.

If I may be of further assistance regarding this matter, please feel free to contact me.

Best wishes,

Erin M. Buck
Assistant to the Publisher
American Mathematical Society
201 Charles Street
Providence, RI 02904

Phone: (401) 455-4141 or 800-321-4267 ext. 4141
Fax: (401) 331-3842

-----Original Message-----
From: banire@mcb.mcgill.ca [mailto:banire@mcb.mcgill.ca]

Sent: Wednesday, April 29, 2009 11:41 AM
To: bookstore@ams.org
Cc: banire@mcb.mcgill.ca; banire@gmail.com
Subject: written authorization to include one of my published paper in my Ph.D. thesis

Dear publishers,
I would like to include my published paper below into my Ph.D. thesis manuscript. For this purpose, I need a permission from the editor. Thus, it will be very helpful to have an authorization from you. You will find below the complete reference of this manuscript.

Thank you and Best regards
Abdoulaye Baniré Diallo

Makarenkov, V., Boc, A., Boubacar Diallo, A., Baniré Diallo, A. (2008) Algorithms for detecting complete and partial horizontal gene transfers: theory and practice. In CRM Proceedings and AMS Lecture Notes (Pardalos, P.
M., Hansen P., eds.), Vol. 45, pp. 159–79.

J'autorise Abdoulaye Baniré Diallo à inclure nos articles en commun dans sa thèse de doctorat. Ces articles sont :

1. Diallo, A.B., Nguyen, D., Badescu, D., Boc, A., Blanchette, M.and Makarenkov, V. (2009): Étude de classification des bactériophages. Mathématiques, Informatique et Sciences Humaines. 16 pages. In preparation.

2. Diallo, A.B., Badescu, D., Makarenkov, V., Blanchette, M. (2009): A whole genome study and identification of specific carcinogenic regions of the Human Papilloma Viruses. Journal of Computational Biology. Accepted for publication.

En foi de quoi, je lui délivre cette attestation pour servir et valoir ce que de droit.

Dunarel Badescu

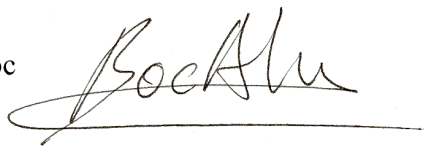
Date : 4 Mai 2009

J'autorise Abdoulaye Baniré Diallo à inclure nos articles en commun dans sa thèse de doctorat. Ces articles sont :

1. Diallo, A.B., Nguyen, D., Badescu, D., Boc, A., Blanchette, M.and Makarenkov, V. (2009): Étude de classification des bactériophages. Mathématiques, Informatique et Sciences Humaines. 16 pages. In preparation.

2. Makarenkov, V., Boc, A., Diallo Al. Bo. and Diallo Ab.Ba. (2008): Algorithms for detecting complete and partial horizontal gene transfers: Theory and practice, in Data Mining and Mathematical Programming, P.M. Pardalos and P. Hansen eds., CRM Proceedings and AMS Lecture Notes, 45, 159-179.

En foi de quoi, je lui délivre cette attestation pour servir et valoir ce que de droit.

Alix Boc 

Date : 04 mai 2009

J'autorise Abdoulaye Baniré Diallo à inclure notre article en commun dans sa thèse de doctorat. Cet article est :

Makarenkov, V., Boc, A., Diallo Al. Bo. and Diallo Ab.Ba. (2008): Algorithms for detecting complete and partial horizontal gene transfers: Theory and practice, in Data Mining and Mathematical Programming, P.M. Pardalos and P. Hansen eds., CRM Proceedings and AMS Lecture Notes, 45, 159-179.

En foi de quoi, je lui délivre cette attestation pour servir et valoir ce que de droit.

Alpha Boubacar Diallo

Date : 04 - 05 - 2009

From: Dung Nguyen <dung2kim@yahoo.ca>
Subject: **Re : autorization to include**
Date: May 4, 2009 2:17:12 PM GMT-04:00
To: Abdoulaye Banire Diallo <banire@gmail.com>

Bonjour Abdoulaye,

Tu as mon consentement.
Bonne chance pour la suite des événements.

Dung Nguyen

De : Abdoulaye Banire Diallo <banire@gmail.com>
À : Dung Nguyen <dung2kim@yahoo.ca>
Envoyé le : lundi 4 mai 2009, 11 h 57 min 43 s
Objet : autorization to include

Bonjour Nguyen,
Il faut que j'aie une autorisation écrite de ta part pour inclure notre article en commun dans ma thèse:

Diallo, A.B., Nguyen, D., Badescu, D., Boc, A., Blanchette, M. and Makarenkov, V. (2009): Étude de classification des bactériophages. Mathématiques, Informatique et Sciences Humaines. 16 pages. In preparation.

Tu pourrais repondre à cet e-mail en m'indiquant ton autorisation. Je dois le faire pour tous les co-auteurs de tous mes articles. Merci pour une reponse rapide. La date limite de dépot étant le 6 mai. Je l'ai su la semaine dernière.
Abdoulaye

--
Abdoulaye Baniré Diallo
Professeur
Département d'informatique
Université du Québec À Montréal
Canada

From: Eric Gaul <ericgaul@yahoo.ca>
Subject: **RE : nouvelles et demande permission**
Date: June 4, 2008 8:26:19 PM GMT-04:00
To: Abdoulaye Banire Diallo <banire@mcb.mcgill.ca>

Bonjour Abdoulaye,

tu as ma permission. Pour le site, je ne crois pas
que ce soit encore actif, par contre.

Cela va très bien à mon collège. On engage toujours
plus de nouveaux enseignants, ce qui fait que je me
sens en sécurité pour mon emploi.

Nous avons eu notre 4ème et dernière enfant à Noël.
La famille est maintenant terminée...

Je souhaite qu'un climat plus propice à la recherche
et à l'enseignement revienne à l'UQAM, pour vous tous !

Éric.

--- Abdoulaye Banire Diallo <banire@mcb.mcgill.ca> a écrit :

> Bonjour Eric,
> Cela fait un bout de temps qu'on s'est pas parlé. Cela me manque de ne plus te voir à l'UQAM, surtout toutes les discussions que nous avions. J'espère que cela se passe bien avec la nouvelle de la famille? Par ailleurs, je t'informe que je dois déposer ma thèse en début juin. Je voulais avoir ta permission pour mettre une partie du travail que nous avions fait ensemble pour la visualisation des indels en 2006 que nous avions pas pu terminer. Il y aura une mention explicite de toi et ton travail et une reference vers ton site web ou est hébergé le prototype si il est toujours actif? Si tu as des questions la dessus n'hesite pas. J'attends une réponse rapide car je n'ai pas commencé encore la partie représentabilité des indels. Merci pour tout.
> Abdoulaye
> PS: commence cela se passe à ton Collège? À l'UQAM actuellement, l'ambiance est morose avec toutes les coupures dans tout.
>
> Abdoulaye