Estimating phylogenies from molecular data

Daniele Catanzaro

Abstract Phylogenetic estimation from aligned DNA, RNA or amino acid sequences has attracted more and more attention in recent years due to its importance in analysis of many fine-scale genetic data. Nowadays, its application fields range from medical research to drug discovery, to epidemiology, to systematics and population dynamics. Estimating phylogenies involves solving an optimization problem, called the *Phylogenetic Estimation Problem* (PEP), whose versions depend on the criterion used to select a phylogeny among plausible alternatives. This chapter offers an overview of PEP and discuss the most important versions that occur in the literature.

1 Introduction

Molecular phylogenetics studies the hierarchical evolutionary relationships among species, or *taxa*, by means of molecular data such as DNA, RNA, amino acid or codon sequences. These relationships are usually described through a weighted tree, called a *phylogeny* (see Fig. 1), whose *leaves* represent the observed taxa, *internal vertices* represent the intermediate ancestors, *edges* represent the estimated evolutionary relationships, and *edge weights* represent measures of the similarity between pairs of taxa.

Phylogenies provide a fundamental information in analysis of many fine-scale genetic data, for this reason the use of molecular phylogenetics has become more and more frequent (and sometimes indispensable) in several research fields such as systematics, medical research, drug discovery, epi-

Dr. Daniele Catanzaro,
Service Graphes and Mathematical Optimization, Computer Science Department, Université Libre de Bruxelles (U.L.B.), Boulevard du Triomphe, CP 210/01, B-1050, Brussels, Belgium. Phone: 0032 2 650 5628. Fax: 0032 2 650 5970. e-mail: dacatanz@ulb.ac.be

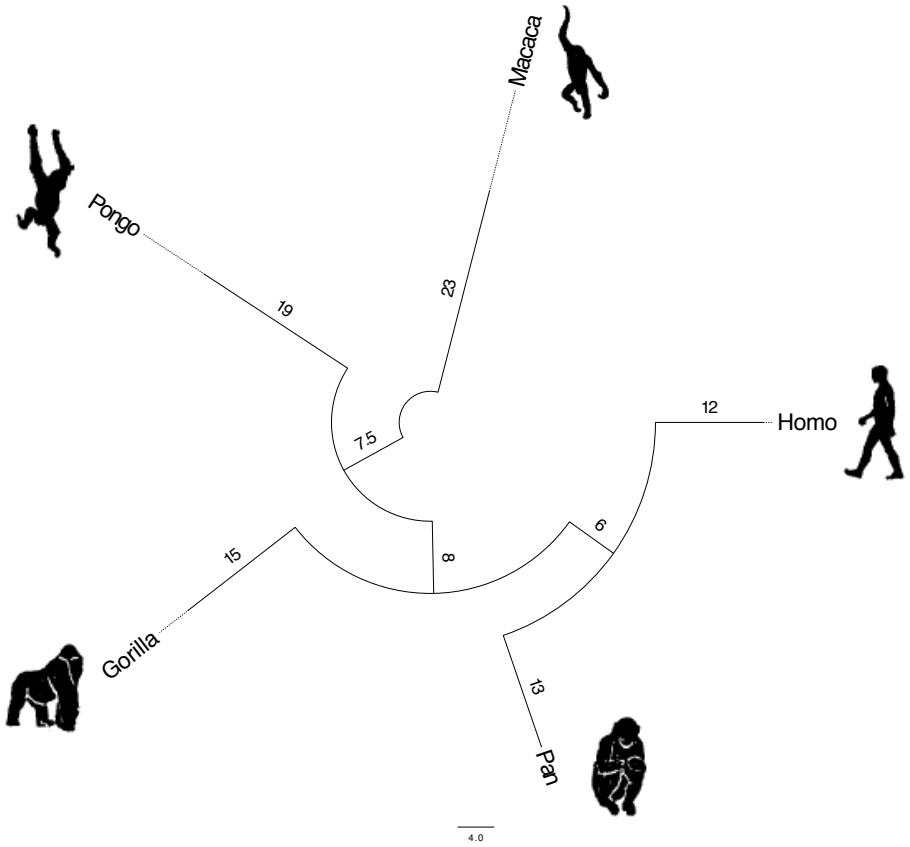


Fig. 1 An example of a phylogeny of primates.

demiology, and population dynamics [55]. For example, the use of molecular phylogenetics was of considerable assistance to predict the evolution of human influenza A [8], to understand the relationships between the virulence and the genetic evolution of HIV [54, 65], to identify emerging viruses as SARS [50], to recreate and investigate ancestral proteins [17], to design neuropeptides causing smooth muscle contraction [2], and to relate geographic patterns to macroevolutionary processes [35].

Since no one could practicably observe evolution over thousands or millions of years, a part from known phylogenies [see 56], there is no general way to validate empirically a candidate phylogeny for a set of molecular sequences extracted from taxa. For this reason, the literature proposes a number of criteria for selecting one phylogeny from among plausible alternatives. Each criterion adopts its own set of evolutionary hypotheses, whose ability to describe evolution of taxa determines the gap between the *real* and the *true*

phylogeny, i.e., the gap between the real evolutionary process of taxa and the phylogeny that one would obtain under the same set of hypotheses if all molecular data from taxa were available [9].

The criteria of phylogenetic estimation can usually be quantified and expressed in terms of objective functions, giving rise to families of optimization problems whose general paradigm can be stated as follows:

Problem 1 – *The Phylogenetic Estimation Problem (PEP)*

$$\begin{aligned} \textit{optimize} \quad & f(T) \\ \textit{s.t.} \quad & g(\Gamma, T) = 0 \\ & T \in \mathcal{T} \end{aligned}$$

where Γ is the set of molecular sequences from n taxa, T a phylogeny of Γ, $\mathcal{T}$ the set of $(2n-5)!! = 1 \times 3 \times 5 \times 7 \cdots \times 2n-5$ phylogenies of Γ, $f : \mathcal{T} \to \mathbb{R}$ a function modeling the selected criterion of phylogenetic estimation, and $g : \Gamma \times \mathcal{T} \to \mathbb{R}$ a function correlating the set Γ to a phylogeny T.

A specific optimization problem, or *phylogenetic estimation paradigm*, is completely characterized by defining the functions f and g. The phylogeny T^* that optimizes f and satisfies g is referred to as *optimal*, and if T^* approaches the true phylogeny as the amount of molecular data from taxa increases, the corresponding criterion is said to be *statistically consistent* [31]. The statistical consistency is a desirable property in molecular phylogenetics because it measures the ability of a criterion to recover the true (and hopefully the real) phylogeny of the given molecular data. Later in the chapter we will show that the consistency property changes from criterion to criterion and in some cases may be even absent.

Here we provide a review of the main estimation criteria that occur in the literature on molecular phylogenetics. Particular emphasis is given to the comparative description of the hypotheses at the core of each criterion and to the optimization aspects related to the phylogenetic estimation paradigms. In Section 2 we discuss the problem of measuring the similarity among molecular sequences. In Section 3 we discuss the fundamental least-squares paradigm and formalize the concept of phylogeny. In Section 4 we present the minimum evolution paradigm, by evidencing the recent perspectives and computational advances. Finally, in Section 5 we present the likelihood and the bayesian paradigms by exposing briefly their benefits and drawbacks.

2 Measuring molecular similarity

The degree of similarity between pairwise molecular sequences reflects the amount of mutation events that occurred since they split from their common ancestor. Quantifying such similarity constitutes the first step in the phylo-

genetic estimation process [11]. The task involves the investigation and the modeling of the *mutation process* over time, i.e., the process by which errors occurs in molecular data and are inherited between generations.

Different types of mutation may occur in the genome structure, most of which are point mutations, i.e., changes that involve the replacement, or *substitution*, of one nucleotide for another in the DNA sequence. Point mutations can be classified in two categories: the transitions and the transversions. The transitions occur when a purine nucleotide (adenine or guanine) is substituted for another purine, or when a pyrimidine (cytosine or thymine) is substituted for another pyrimidine. The transversions occur when a pyrimidine is substituted for a purine, or vice versa.

A second class of point mutations are those that lead to *insertions* and *deletions* of nucleotides in the genome. This phenomenon mainly occurs in non-coding regions of DNA, but may interest also coding regions of the genome and be the cause of deleterious effects [see 56].

Finally a third class of mutations are those that involve entire chromosome regions of the genome. Specifically, we may have: (i) a *duplication*, when a chromosome region is duplicated; (ii) a *translocation*, when a chromosome region is transferred into another chromosome; (iii) an *inversion*, when a chromosome region is broken off, turned upside down and reconnected; (iv) a *deletion*, when a chromosome region is missing or deleted; (v) and a *loss of heterozygosity*, e.g., when two instances of the same chromosome break and then reconnect but to the different end pieces [see 56].

Modeling the second and the third classes of mutations is generally non-trivial and requires advanced mathematical background. We refer the interested reader to Felsenstein [28] for an introduction and to Park and Deem [57] for recent advances in the modeling of such classes. Here we shall focus on the first class of mutations and present a fundamental model of molecular evolution which is at the core of the most currently used criteria of phylogenetic estimation. Unless not stated otherwise, throughout the chapter we will always assume that the molecular sequences under study have been previously subjected to an *alignment process*, i.e., a process through which the evolutionary relationships between nucleotides of molecular data are evidenced [see 59, for details].

2.1 The time homogeneous Markov model of molecular evolution

Let S be a DNA sequence, i.e., a string of fixed length over an alphabet $\Upsilon = \{A, C, G, T\}$, where 'A' codes for Adenine, 'C' for Cytosine, 'G' for Guanine, and 'T' for Thymine. Let $r_{ij} \geq 0$, $i \neq j$, be the constant rate of substitution from nucleotide i to nucleotide j. Assume that each character (site) of S evolves independently over time and that, instant per instant, the

Markov conservative hypothesis [38] holds, i.e.,

$$r_{ii} = -\sum_{j\in\Upsilon,\ j\neq i} r_{ij} \quad \forall\ i \in \Upsilon. \tag{1}$$

Let $p_{ij}(t)$ be the probability that nucleotide i undergoes to a substitution to nucleotide j at finite time t. Then, if the superposition principle holds, at $t + dt$ such probability can be written as

$$p_{ij}(t + dt) = \sum_{k\in\Upsilon} p_{ik}(t)p_{kj}(dt) \quad \forall\ i, j \in \Upsilon. \tag{2}$$

By subtracting $p_{ij}(t)$ in both sides of equation (2) and dividing for dt we obtain:

$$\frac{p_{ij}(t + dt) - p_{ij}(t)}{dt} = \frac{\sum_{k\in\Upsilon,\ k\neq j} p_{ik}(t)p_{kj}(dt)}{dt} + p_{ij}(t)\frac{p_{jj}(dt) - 1}{dt} \quad \forall\ i, j \in \Upsilon,$$

i.e.,

$$\begin{aligned}\frac{p_{ij}(t + dt) - p_{ij}(t)}{dt} &= \frac{\sum_{k\in\Upsilon,\ k\neq j} p_{ik}(t)p_{kj}(dt)}{dt} \\ &\quad + p_{ij}(t)\frac{1 - \sum_{k\in\Upsilon,\ k\neq j} p_{kj}(dt) - 1}{dt} \quad \forall\ i, j \in \Upsilon.\end{aligned}$$

Hence we have

$$\dot{p}_{ij}(t) = \sum_{k\in\Upsilon,\ k\neq j} p_{ik}(t)r_{kj} + p_{ij}(t)r_{jj} \quad \forall\ i, j \in \Upsilon. \tag{3}$$

When expressing equation (3) in matrix form, the Chapman-Kolmogorov master equation arises

$$\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{R} = \mathbf{R}\mathbf{P}(t)$$

whose integral

$$\mathbf{P}(t) = \mathbf{e}^{\mathbf{R}t} = \sum_{n=0}^{\infty} \frac{\mathbf{R}^n t^n}{n!} \tag{4}$$

is known as the *Time Homogeneous Markov* (THM) model of DNA sequence evolution [47, 62]. The THM model is a generalization of the Markov models described in Jukes and Cantor [43], Kimura [45], Hasegawa et al. [36], Tamura and Nei [77], and can be easily adapted to RNA, amino acid and codon sequences as shown in Felsenstein [28] and Schadt and Lange [70, 71]. In the next section we shall investigate the dynamics of the THM model in order to

derive a commonly used formula to quantify the similarity between molecular data.

2.2 Estimating evolutionary distances from molecular data

Two molecular sequences S_1 and S_2, evolving at time t_0 from a common ancestor, could be characterized at time t by different amounts of substitution events, some of which not directly observable. Hence, if we would sample the sequences at time t and measure their similarity, or *evolutionary distance*, in terms of number of observed differences, we could underestimate the overall substitution events that occurred since S_1 and S_2 split from their common ancestor. A number of authors suggested that the use of the time homogeneous Markov models could overcome the underestimation problem in all those cases in which the hypotheses at the core of the model would properly describe the real evolutionary process of the analyzed sequences [28]. Moreover, in order to compare the evolutionary distances of different pairs of molecular sequences, the authors also proposed to express the evolutionary distances in terms of expected number of substitution events per site rather than the time necessary to transform a sequence into another [28]. In this section we will present the most general formula currently known in the literature to compute the evolutionary distance from pairwise molecular sequences. To this aim, we shall investigate now the dynamics of the THM model.

As shown in Zadeh and Desoer [83], equation (4) can also be expressed in closed formula as

$$\mathbf{P}(t) = \mathbf{e}^{\mathbf{R}t} = \Omega \mathbf{e}^{\Lambda t} \Omega^{-1}, \tag{5}$$

where Ω is the eigenvector matrix of $\mathbf{R}$, and Λ is the diagonal matrix of the eigenvalues of $\mathbf{R}$ respectively. This fact suggests that the spectrum of $\mathbf{P}(t)$ is the exponential spectrum of $\mathbf{R}$, i.e., that the dynamics of $\mathbf{P}(t)$ is univocally determined from the knowledge of the spectrum of $\mathbf{R}$ [83].

It is worth noting that the Markov conservative hypothesis implies that the determinant of matrix $\mathbf{R}$ is equal to zero, i.e., at least one of its eigenvalues is identically zero. Moreover, since any k-leading principal sub-matrix of $\mathbf{R}$, $k < 4$, has negative determinant, for one of the Sylvester corollaries [see 6, p. 409] all the remaining eigenvalues are negative. Thus, as the spectrum of $\mathbf{P}(t)$ is the exponential spectrum of $\mathbf{R}$, matrix $\mathbf{P}(t)$ has at least one eigenvalue equal to 1, called *the maximal Lyapunov exponent*, and three eigenvalues lying in the interval $[0, 1]$. The maximal Lyapunov exponent prevents the presence of chaotic attractors and guarantees that, as t goes to infinity, the generic entry $p_{ij}(t)$ is non-zero and independent on the starting state $i \in \Upsilon$. In

other words, the maximal Lyapunov exponent guarantees the existence of four positive values π_A, π_C, π_G, and π_T, called *equilibrium frequencies*, such that

$$\lim_{t\to\infty} p_{ij}(t) = \pi_j \qquad \forall\ i,j \in \Upsilon.$$

The values π_j constitute a *stationary distribution* and turn out useful to measure the evolutionary distance between S_1 and S_2. In fact, denote $\mathbf{O}(t)$ as a matrix whose generic entry $o_{ij}(t)$, $i,j \in \Upsilon$, represents the probability that, at a given site and time t, S_1 is characterized by nucleotide i and S_2 by nucleotide j. Assume that $\mathbf{O}(0) = \Pi$, where Π denotes a diagonal matrix whose j-th diagonal entry is π_j. Then it holds that

$$o_{ij}(t) = \sum_{k\in\Upsilon} p'_{ik}(t)\pi_k p_{kj}(t) \qquad \forall\ i,j \in \Upsilon,\ t \geq 0,$$

or equivalently

$$\mathbf{O}(t) = \mathbf{P}'(t)\Pi\mathbf{P}(t) \qquad t \geq 0 \tag{6}$$

where $\mathbf{P}'(t)$ denotes the transpose of $\mathbf{P}(t)$. Premultiplying for Π^{-1} both sides of equation (6) we have

$$\Pi^{-1}\mathbf{O}(t) = \Pi^{-1}\mathbf{P}'(t)\Pi\mathbf{P}(t) = \Pi^{-1}\mathbf{e}^{\mathbf{R}'t}\Pi\mathbf{e}^{\mathbf{R}t}.$$

Since for any matrix function $f(\mathbf{ABA}^{-1}) = \mathbf{A}f(\mathbf{B})\mathbf{A}^{-1}$, we have

$$\Pi^{-1}\mathbf{O}(t) = \mathbf{e}^{\Pi^{-1}\mathbf{R}'t\Pi}\mathbf{e}^{\mathbf{R}t}. \tag{7}$$

If we assume that the hypothesis of *time-reversibility* holds, i.e., that

$$\Pi\mathbf{R} = \mathbf{R}'\Pi,$$

then $\Pi^{-1}\mathbf{R}'t\Pi$ and $\mathbf{R}t$ are commutative, and equation (7) becomes

$$\Pi^{-1}\mathbf{O}(t) = \mathbf{e}^{\Pi^{-1}\mathbf{R}'t\Pi+\mathbf{R}t}. \tag{8}$$

By applying the logarithmic matrix function to both members of equation (8) and premultiplying for Π we obtain

$$\mathbf{R}'t\Pi + \Pi\mathbf{R}t = \Pi\log(\Pi^{-1}\mathbf{O}(t)).$$

As the negative trace of $2t\Pi\mathbf{R}$ represents the expected number of substitution events per site between S_1 and S_2, at time t the evolutionary distance d_{S_1,S_2} between S_1 and S_2 can be computed as

$$d_{S_1,S_2} = -2t\ \mathrm{tr}[\Pi\mathbf{R}] = -\mathrm{tr}[\Pi\log(\Pi^{-1}\mathbf{O}(t))]. \tag{9}$$

Equation (9) is known as the *General Time-Reversible* (GTR) distance [47, 62], and is the most general formula to quantify the similarity between molecular data by using a time-reversible Markov model of molecular evolution. It is worth noting that, if from one hand the hypothesis of time-reversibility simplifies the formalization of the evolutionary process of a pair of molecular sequences, from the other its introduction gives rises to important consequences. In fact, the hypothesis of time-reversibility implies that if we would compare two molecular data whose nucleotide frequencies are in equilibrium, the probability that a nucleotide i undergoes to a substitution to nucleotide j would be equal to the probability that a nucleotide j undergoes to a substitution to nucleotide i. Thus, given a present-day molecular sequence and its ancestral sequence, it would not possible to determine which sequence is the present and which is the ancestral one. Hence, the hypothesis of time-reversibility removes the temporal directionality from the evolutionary process. We shall show in the next sections how the paradigms of phylogenetic estimation take advantage of this fact. Below we provide an example from Waddell and Steel [78] showing a possible application of (9).

2.2.1 Estimating evolutionary distances from molecular data: A practical example

Consider the mitochondrial DNA sequences of human and chimpanzee showed in Horai et al. [39]. The corresponding matrices $\mathbf{O}(t)$ and Π are respectively

$$\mathbf{O}(t) = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \left(\begin{array}{cccc} 0.2889 & 0.0012 & 0.0131 & 0.0005 \\ 0.0012 & 0.2799 & 0.0001 & 0.0266 \\ 0.0131 & 0.0001 & 0.1180 & 0.0001 \\ 0.00005 & 0.0266 & 0.0001 & 0.2299 \end{array} \right) \end{array} \begin{array}{c} A \\ C \\ G \\ T \end{array}$$

and

$$\Pi = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \left(\begin{array}{cccc} 0.3037 & 0 & 0 & 0 \\ 0 & 0.3079 & 0 & 0 \\ 0 & 0 & 0.1313 & 0 \\ 0 & 0 & 0 & 0.2571 \end{array} \right) \end{array} \begin{array}{c} A \\ C \\ G \\ T. \end{array}$$

The product $\Pi^{-1}\mathbf{O}(t)$ is

$$\Pi^{-1}\mathbf{O}(t) = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \left(\begin{array}{cccc} 0.9513 & 0.0040 & 0.0430 & 0.0017 \\ 0.0040 & 0.9092 & 0.0003 & 0.0865 \\ 0.0995 & 0.0008 & 0.8989 & 0.0008 \\ 0.0030 & 0.1036 & 0.0004 & 0.8940 \end{array} \right) \end{array} \begin{array}{c} A \\ C \\ G \\ T \end{array}$$

and the corresponding logarithm matrix function $\log(\Pi^{-1}\mathbf{O}(t))$ is

$$\Pi^{-1}\mathbf{O}(t) = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -0.0524 & 0.0042 & 0.0466 & 0.0016 \\ 0.0042 & -0.1008 & 0.0002 & 0.0963 \\ 0.1078 & 0.0006 & -0.1091 & 0.0007 \\ 0.00019 & 0.1154 & 0.0004 & -0.1176 \end{pmatrix} \end{array} \begin{array}{c} A \\ C \\ G \\ T. \end{array}$$

The product $\Pi \log(\Pi^{-1}\mathbf{O}(t))$ is

$$\Pi \log(\Pi^{-1}\mathbf{O}(t)) = \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -0.0159 & 0.0013 & 0.0142 & 0.0005 \\ 0.0013 & -0.0310 & 0.0001 & 0.0297 \\ 0.0142 & 0.0001 & -0.0143 & 0.0001 \\ 0.0005 & 0.0293 & 0.0001 & -0.0302 \end{pmatrix} \end{array} \begin{array}{c} A \\ C \\ G \\ T \end{array}$$

whose negative trace provides the evolutionary distance $d = -\mathrm{tr}[\Pi \log(\Pi^{-1}\mathbf{O}(t))] = 0.09152$.

The reader interested in more sophisticated applications of the GTR distance will find useful examples in Lanave et al. [47], Rodriguez et al. [62], and Catanzaro et al. [10, 11].

3 The Least-Squares paradigm of phylogenetic estimation

A paradigm of phylogenetic estimation is a quantitative criterion used to discern a phylogeny from among plausible alternatives. One of the earliest paradigms was introduced by Cavalli-Sforza and Edwards [15] and is known as the *additive model* or the *the least-squares model* of phylogenetic estimation [9].

Cavalli-Sforza and Edwards observed that, as molecular data provide the most detailed anatomy possible for any organism, the diversity of life on Earth must be reflected in them. Hence, if evolution of a set of molecular data from taxa could be seen as a tree, then it could be described through a process that changes nucleotides over time. The trajectories described by such a process would split as taxa diverges, unite as taxa hybridize, end as taxa become extinct, and living taxa would be represented by the intercept of the process and the "now" plane (see Fig. 2).

In general, we do not have a sampling of such a process over time but only the knowledge of the living taxa. Hence, in absence of further information, one may be able only to reconstruct the projection of the process onto the "now" plane rather than the process itself. Note that, while the evolutionary process over time is "directed", its projection is not (see Fig. 2). Thus, when

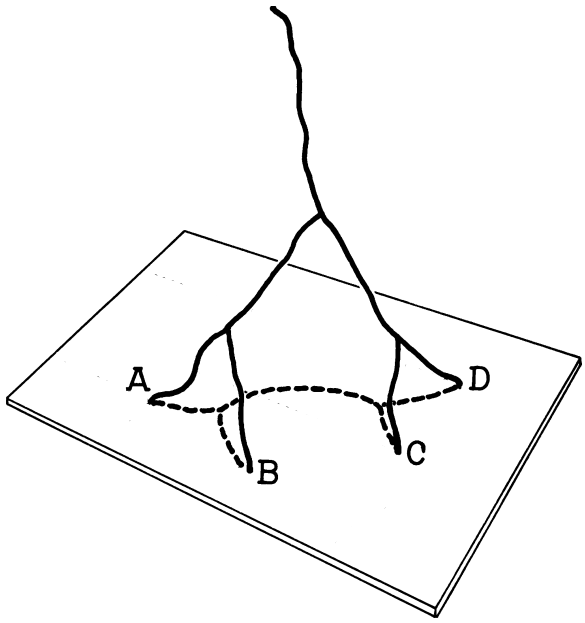


Fig. 2 An evolutionary process and its projection onto the "now" plane - from Cavalli-Sforza and Edwards [15].

the projection is considered, the direction of evolution is definitely missed. Nevertheless, the projection of the evolutionary process constitutes still an important piece of information for the analyzed taxa, for this reason Cavalli-Sforza and Edwards proposed a possible paradigm to recover it.

The authors first considered the problem of how to represent formally a projection (phylogeny) of the evolutionary process. In order to remark the lack of a direction in evolution, the authors proposed to remove the root and the orientation in the edges of a phylogeny, and represented it as an unrooted binary tree, i.e., an undirected acyclic graph in which each internal vertex has degree three. The degree constraint has not necessarily a biological foundation but helped the authors to formalize the evolutionary process. In fact, given n taxa, the degree constraint implies that the number of edges in a phylogeny T is $(2n-3)$ and the number of internal vertices is $(n-2)$. To prove the claim note that, as T is a tree, it holds that

$$|\mathcal{E}_i(T)| + |\mathcal{E}_e(T)| = |V_i| + |V_e| - 1, \tag{10}$$

where $\mathcal{E}_e(T)$ and $\mathcal{E}_i(T)$ are the set of external and internal edges of T, respectively. Moreover, since internal vertices have degree three, the following property holds:

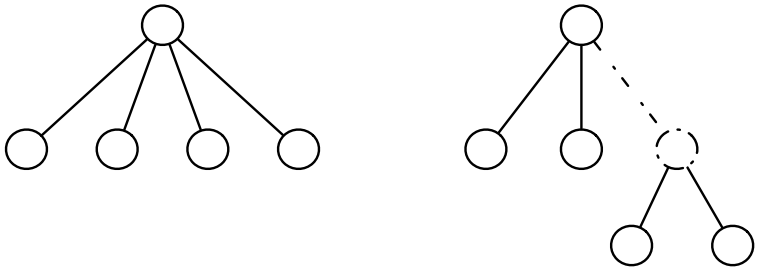


Fig. 3 The 4-ary tree (on the left) can be transformed into an unrooted binary tree by adding a dummy vertex and edge (dashed, on the right).

$$2|\mathcal{E}_i(T)| + 2|\mathcal{E}_e(T)| = 3|V_i| + |V_e|. \tag{11}$$

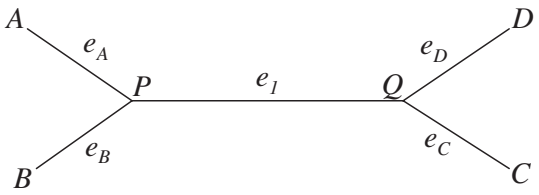
Combining (10) and (11) it follows that $|V_i| = (n-2)$ and $|\mathcal{E}_i| = (n-3)$. Thus, a phylogeny $T \in \mathcal{T}$ can be seen as an unrooted binary tree in which the n taxa are the n leaves of T and the common ancestors are internal vertices of degree three. It is worth noting that dealing with unrooted binary trees does not introduces oversimplifications since it is easy to see that any m-ary tree can be transformed into a phylogeny by adding "dummy" vertices and edges (e.g., see Fig. 3).

Cavalli-Sforza and Edwards encoded a phylogeny in $\mathcal{T}$ by means of an *Edge–Path incidence matrix of a Tree* (EPT) [see 52, p. 550] i.e., a network matrix $\mathbf{X}$ having a row for each path between two leaves and a column for each edge. The generic entry $x_{rs,e}$ of matrix $\mathbf{X}$ is equal to 1 if edge e belongs to the path p_{rs} from leaf r to leaf s and 0 otherwise. As an example, Fig. 4(b) shows the EPT matrix corresponding to the phylogeny shown in Fig. 4(a). Hence, the authors proposed a model in which each evolutionary distance d_{rs}, $r, s \in \Gamma$, among pairwise molecular data could be thought of as the resulting sum of mutation events accumulated on edges belonging to the path p_{rs} linking taxa r and s on $\mathbf{X}$. In other words, fixed a phylogeny $\mathbf{X}$ and defined w_e as the amount of mutation events on edge e, Cavalli-Sforza and Edwards asserted that

$$\mathbf{X}\mathbf{w} = \mathbf{D}^{\triangle} \tag{12}$$

where $\mathbf{w} = \{w_e\}$ is the edge weight vector associated to $\mathbf{X}$, and $\mathbf{D}^{\triangle}$ is a $n(n-1)/2$ vector whose components are obtained by taking row by row the entries of the strictly upper triangular matrix $\mathbf{D} = \{d_{rs}\}$.

In general, for a fixed matrix $\mathbf{X}$, equation (12) may not admit solutions, for this reason the authors proposed the use of the *Ordinary Least-Squares* (OLS) to find the entries of vector $\mathbf{w}$. Specifically, the authors suggested that the values $\rho_{rs} = \sum_{e \in p_{rs}} x_{rs,e} w_e$ should minimize the function



(a)

	e_A	e_B	e_C	e_D	e_1
Path AB	1	1	0	0	0
Path AC	1	0	1	0	1
Path AD	1	0	0	1	1
Path BC	0	1	1	0	1
Path BD	0	1	0	1	1
Path CD	0	0	1	1	0

(b)

Fig. 4 (a) An example of a phylogeny of four taxa (modeled as an unrooted binary tree in which each internal vertex has degree three) and its associated EPT matrix (b).

$$\sum_{r,s\in\Gamma:r\neq s}(d_{rs}-\rho_{rs})^2=\sum_{r,s\in\Gamma:r\neq s}(d_{rs}-\sum_{e\in p_{rs}}x_{rs,e}w_e)^2,$$

i.e., minimize the quadratic error related to the approximation of the evolutionary process with its projection. This condition holds when

$$\mathbf{w}=\mathbf{X}^{\dagger}\mathbf{D}^{\triangle},$$

where $\mathbf{X}^{\dagger}$ is the Moore-Penrose pseudo-inverse matrix of $\mathbf{X}$. Thus, Cavalli-Sforza and Edwards' paradigm of phylogenetic estimation may be stated in terms of the following NP-hard convex optimization problem [22]:

Problem 2 – *The Ordinary Least-Squares Problem (OLSP)*

$$\min_{\mathbf{X}\in\mathcal{X},\mathbf{w}\in\mathbb{R}^{2n-3}}\quad f(\mathbf{X})=\sum_{r,s\in\Gamma:r\neq s}(d_{rs}-\sum_{e\in p_{rs}}x_{rs,e}w_e)^2$$

where $\mathcal{X}$ denotes the set of all possible EPT matrices coding phylogenies. We refer the reader interested in a mathematical description of the necessary and sufficient conditions that characterize the set $\mathcal{X}$ to Catanzaro et al. [14].

3.1 Modified least-squares paradigms of phylogenetic estimation

A number of authors proposed some modifications to Cavalli-Sforza and Edwards' model. Specifically, Fitch and Margoliash [30] observed that OLSP implicitly considers the evolutionary distances d_{rs} among pairwise molecular data as uniformly distributed independent random variables, a hypothesis that cannot be considered generally true due to the common evolutionary history of the analyzed taxa and the presence of sampling errors in molecular data. Hence, Fitch and Margoliash proposed to modify Cavalli-Sforza and Edwards' paradigm by introducing the quantities ω_{rs} representing the variances of d_{rs}. They set $\omega_{rs} = 1/d_{rs}^2$, $r, s \in \Gamma$, and stated the following paradigm:

Problem 3 – *The Weighted Least-Squares Problem (WLSP)*

$$\min_{\mathbf{X}\in\mathcal{X},\mathbf{w}\in\mathbb{R}^{2n-3}} \quad f(\mathbf{X}) = \sum_{r,s\in\Gamma:r\neq s} \omega_{rs}(d_{rs} - \sum_{e\in p_{rs}} x_{rs,e}w_e)^2.$$

Later, Chakraborty [16] and Hasegawa et al. [37] proposed a very similar paradigm, called the *Generalized Least-Squares Problem* (GLSP), in which the variances ω_{rs} are replaced by the covariances of d_{rs}. Nowadays, GLSP has fallen into disuse due to its statistical inconsistency problems [9].

3.2 Drawbacks of the least-squares paradigms of phylogenetic estimation

Although the least-squares paradigm is a milestone in molecular phylogenetics, it is characterized by a number of drawbacks. For example, Cavalli-Sforza and Edwards' paradigm returns a *tree metric*, i.e., a phylogeny whose edge weights are non-negative [72, 79], whenever the distance matrix $\mathbf{D}$ satisfies the *ultrametric property*

$$d_{rs} \leq \max\{d_{rq}, d_{qs}\} \qquad r, s, q \in \Gamma \ : \ r \neq s \neq q$$

or the *additive property*

$$d_{rs} + d_{hk} \leq \max\{d_{rh} + d_{sk}, d_{rk} + d_{sh}\} \qquad r, s, h, k \in \Gamma \ : \ r \neq s \neq h \neq k.$$

Specifically, when $\mathbf{D}$ is ultrametric or additive the solution of Problem 2 is unique and obtainable in polynomial time through the UPGMA greedy algorithm [73] or the Sequential algorithm [79], respectively.

Unfortunately, when $\mathbf{D}$ is generic (e.g., when it is obtained by means of the THM model, see Section 2), the least-squares paradigm may lead to the

occurrence of negative entries in the vector $\mathbf{w}$, i.e., to a phylogeny that is not a tree metric [31, 46]. Negative edge weights are infeasible both from a conceptual point of view (a distance, being an expected number of mutation events over time, cannot be negative [44]) and from a biological point of view (evolution cannot proceed backwards [56, 76]). For the latter reason at least, non-tree metric phylogenies are generally not accepted in molecular phylogenetics [34].

In response, some authors investigated the consequences of adding or guaranteeing the positivity constraint of edge weights in the least-squares paradigm.

Gascuel and Levy [32] observed that the presence of the positivity constraint transforms any least-square model into a non-negative linear regression problem which involves projecting the distance matrix $\mathbf{D}$ onto the positive cone defined by the set of tree metrics [see also 5, p. 187]. Thus, the authors designed an iterative polynomial time algorithm able to generate a sequence of least-squares projections of $\mathbf{D}$ onto such a set until an additive distance matrix (and the corresponding phylogeny) is obtained.

Farach et al. [26] proposed an alternative approach to impose the positivity constraint. Specifically, the authors proposed to find the minimal perturbation of the distance matrix $\mathbf{D}$ that guarantees the satisfaction of the additive or the ultrametric property. Farach et al. [26] proposed the $\mathcal{L}_\infty$-norm and $\mathcal{L}_1$-norm to constraint the entries of $\mathbf{D}$ to satisfy the additive (ultrametric) property, and proved that such a problem can be solved in polynomial time when $\mathbf{D}$ is required to be ultrametric under the $\mathcal{L}_\infty$-norm. By contrast, the authors proved that their approaches become hard when an ultrametric or an additive distance matrix is required under the $\mathcal{L}_1$-norm.

Finally, Barthélemy and Guénoche [3] and Makarenkov and Leclerc [49] proposed a Lagrangian relaxation of the positivity constraint to guarantee metric trees. Both algorithms are iterative and apply to the OLSP and the WLSP, respectively. Specifically, starting from a leaf, the algorithms generate a phylogeny with a growing number of leaves by solving an optimization problem in which the best non-negative edge weights that minimize the OLSP (respectively the WLSP) are found. Both algorithms are polynomial time and characterized by a computational complexity of $O(n^4)$ and $O(n^5)$, respectively.

A second and possibly more serious drawback of the least-squares is the statistical inconsistency of some paradigms. Specifically, a part from the OLSP which proves to be statistically consistent [23, 67], the only case in which the WLSP is known to be consistent is when the variances ω_{rs} are set to the inverse of the product of two strictly positive constants α_i and α_j. By contrast the GLSP is generally inconsistent [34].

4 The minimum evolution paradigm of phylogenetic estimation

Kidd and Sgaramella-Zonta [44] and Beyer et al. [4] independently proposed an alternative paradigm known as the *minimum evolution problem* or *the minimum evolution paradigm* of phylogenetic estimation [9].

The minimum evolution paradigm arises from Cavalli-Sforza and Edwards' model but mainly differs for the way in which a phylogeny is chosen from among possible alternatives. In fact, the minimum evolution criterion states that, if the evolutionary distances d_{rs} were unbiased estimates of the *true evolutionary distances* (i.e., the distances that one would obtain if all the molecular data from the analyzed taxa were available), then the true phylogeny would have an expected length shorter than any other possible phylogeny compatible with $\mathbf{D}$. Hence, the minimum evolution paradigm aims at finding the phylogeny whose sum of edge weights, estimated from the corresponding evolutionary distances, is minimum [9].

It is worth noting that the minimum evolution criterion does not asses that molecular evolution follows minimum paths, but states, according to classical evolutionary theory, that a minimum length phylogeny may properly approximate the real phylogeny of well-conserved molecular data i.e., data whose basic biochemical function has undergone small change throughout the evolution of the observed taxa [4]. That evolution proceeds by small rather than smallest changes is due to the fact that the neighborhood of possible alleles that are selected at each instant of the life of a taxon is finite, and perhaps more important, the selective forces acting on the taxon may not be constant throughout its evolution [4, 79]. Over the long term (periods of environmental change, including the intracellular environment), small changes will not generally provide the smallest change. Thus, a minimum length phylogeny provides a lower bound on the total number of mutation events that could have occurred along evolution of the observed taxa.

Different versions of the minimum evolution paradigm are discussed in the literature on phylogenetics, and each one is characterized by its own edge weight estimation model [9]. Specifically, we can distinguish between the least-squares edge weight estimation model [24, 67, 68] and the linear programming edge weight estimation model [4, 14, 79]. In the next sections we shall analyze both families in detail.

4.1 The minimum evolution paradigm under the least-squares edge weight estimation model

The earliest minimum evolution paradigm of phylogenetic estimation was proposed by Kidd and Sgaramella-Zonta [44] and exploits Cavalli-Sforza and

Edwards' model to estimate edge weights. The authors proposed to change the objective function of the OLSP with

$$f(\mathbf{X}) = \| \mathbf{w} \|_1 = \| \mathbf{X}^{\dagger}\mathbf{D}^{\triangle} \|_1 \tag{13}$$

giving rise to the following NP-hard convex optimization problem [9]:

Problem 4 – *The Minimum Evolution under Least-Squares Problem (MELSP)*

$$\min_{\mathbf{X}\in\mathcal{X}} \quad f(\mathbf{X}) = \| \mathbf{X}^{\dagger}\mathbf{D}^{\triangle} \|_1 \ .$$

Rzhetsky and Nei [67, 68] observed that the MELSP is statistically consistent, and such a property is also guaranteed when considering a relaxed version of the objective function in which edge weights are summed regardless their sign. However, Swofford et al. [76] criticized the choice of taking into account negative edge weights (or even their absolute value) in the objective function due to their biological unfeasibility. Thus, the authors proposed to replace the objective function (13) with

$$f(\mathbf{X}) = \sum_{e\in\mathcal{E}(T=\mathbf{X})|w_e\geq 0} w_e.$$

Gascuel et al. [34] investigated the statistical consistency of Swofford et al. [76] paradigm, and obtained analogous results to Rzhetsky and Nei [67, 68]. At present, Swofford et al. [76] paradigm is one of the most used versions of minimum evolution, being implemented in the well-known software for phylogenetic estimation "PAUP" [75]. The software is able to solve exactly instances of the paradigm containing up 13 taxa, and implements a hill-climbing metaheuristic to tackle larger instances of the problem.

Recently, Desper and Gascuel [24, 25] formalized the most recent version of the minimum evolution paradigm, called the *Balanced Minimum Evolution Problem* (BME). The paradigm is based on Pauplin [58] seminal work in which the author criticized the biological consideration at the core of the OLSP. In fact, Pauplin noted that, when computing the Moore-Penrose pseudo-inverse of the EPT matrix $\mathbf{X}$, some edges can be weighted more than others. Since there is no biological justification for that, Pauplin proposed a new paradigm in which all edges of a phylogeny were weighted in the same way. The resulting objective function does not depends explicitly on edge weights and can be stated as follows:

$$f(T) = \sum_{r,s\in\Gamma:r\neq s} \frac{d_{rs}}{2^{\tau_{rs}}}$$

where τ_{rs} is called the *topological distance* and denotes the number of edges belonging to the path between taxa r and s in a phylogeny T [9]. Hence, BME can be stated in terms of the following optimization problem:

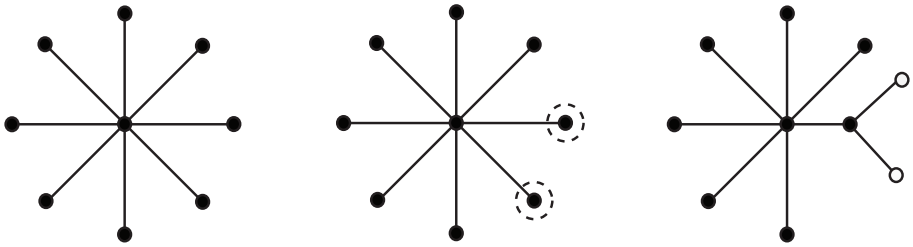


Fig. 5 Clustering heuristics: initially a graph-star is considered; subsequently two vertices (circled) are selected, marked (white vertices) and joined by an internal vertex. The algorithm is iterated on the remaining black vertices until a phylogeny is obtained.

Problem 5 – *The Balanced Minimum Evolution Problem (BME)*

$$\min_{T \in \mathcal{T}} \quad f(T) = \sum_{r,s \in \Gamma : r \neq s} \frac{d_{rs}}{2^{\tau_{rs}}}.$$

BME is known to be statistically consistent [24, 25] and its optimal solution satisfies the positivity constraint whenever the distance matrix satisfies the triangular inequality

$$d_{rs} \leq d_{rq} + d_{qs} \; \forall \; r, s, q \in \Gamma \; : \; r \neq s \neq q.$$

For the latter reason at least, finding the optimal solution to instances of BME is highly desirable. Unfortunately, this task seems hard, although at present no information about the complexity of BME is known in the literature.

Recent advances on the polyhedral combinatorics of BME led to solve exactly instances containing up to 20-25 taxa [13]. However, the size of the instances analyzable to the optimum is still far away from real needs, for this reason it is common the use of clustering heuristics (see Fig. 5), such as the *Neighbor-Joining* Tree (NJT) [see 69, 74], to tackle large instances of BME. Possibly, future developments on the polyhedral combinatorics of BME will provide fundamental new insights for the development of more efficient exact approaches to solution of the problem.

4.2 The minimum evolution paradigm under the linear programming edge weight estimation model

An alternative model to estimate edge weights in the minimum evolution paradigm is provided by linear programming. The model was introduced by

Beyer et al. [4] and is based on the following motivation: if the evolutionary distances between pairs of molecular data have to reflect the number of mutation events required to convert one molecular sequence into another over time, then they must satisfy the triangle inequality. Moreover, since any edge weight of a phylogeny is de facto an evolutionary distance, also the entries of vector $\mathbf{w}$ must satisfy the triangle inequality. This last observation imposes that, for each path p_{rs} from taxa r and s in $\mathbf{X}$, the constraint $\sum_{e \in p_{rs}} w_e x_{rs,e} \geq d_{rs}$ is satisfied. Hence, Beyer et al. [4] proposed a possible paradigm of phylogenetic estimation consisting of solving the following mixed integer programming model:

Problem 6 – *The Minimum Evolution Problem under Linear Programming (MELP)*

$$\begin{aligned} \min_{\mathbf{X} \in \mathcal{X}, \mathbf{w} \in \mathbb{R}_{0^+}^{2n-3}} \quad & f(\mathbf{X}, \mathbf{w}) = \| \mathbf{w} \|_1 \\ s.t. \quad & \mathbf{X}\mathbf{w} \geq \mathbf{D}^{\triangle}. \end{aligned}$$

MELP is a well-known APX-hard problem [26] for which the current exact algorithms described in the literature provide solutions to instances containing no more than a dozen taxa [14]. To the best of our knowledge, nothing is known about the statistical consistency of MELP.

4.3 Drawbacks of the minimum evolution paradigm of phylogenetic estimation

There are mainly two drawbacks that affect the minimum evolution paradigm of phylogenetic estimation: the "rigidity" of its criterion and the hardness of its paradigms.

As regards to the first drawback, some authors, among which notably Felsenstein [28], p. 175, argued that the minimum evolution paradigms could prove unreliable as they neglect rate variation when estimating edge weights. This major criticism could be possibly overcome by using non-homogeneous Markov models. Specifically, In a non-homogeneous Markov model the Chapman-Kolmogorov master equation becomes [83]:

$$\dot{\mathbf{P}}(0,t) = \mathbf{R}(t)\mathbf{P}(0,t), \tag{14}$$

whose integral is given by

$$\mathbf{P}(0,t) = \mathbf{I} + \int_0^t \mathbf{R}(\tau)\mathbf{P}(0,\tau)d\tau. \tag{15}$$

where $\mathbf{I}$ denotes the identity matrix. The use of the integral (15) could prove unpractical for an empirical use. However note that (15) can be approximated through the Peano-Baker sequence

$$\begin{aligned} &\mathbf{P}_0(0,t) = \mathbf{I} \\ &\mathbf{P}_k(0,t) = \mathbf{I} + \int_0^t \mathbf{R}(\tau)\mathbf{P}_{k-1}(0,\tau)d\tau, \quad k = 1,2,\ldots \end{aligned} \tag{16}$$

since it is possible to prove that equation (16) converges to matrix $\mathbf{P}(0,t)$ when $k \to \infty$ [18]. Hence, under a non-homogeneous Markov model, the substitution probability matrix could be easily computed by means of iterative procedures that appropriately approximate (15).

Concerning the second drawback, it is easy to realize that the NP-hardness of the minimum evolution paradigms constitutes a big handicap for the development of exact solution approaches of practical use. Exact approaches are necessary to guarantee the optimality of a given solution and fundamental to investigate whether the hypotheses at the core of a criterion are well suited to describe the evolutionary process of the observed taxa. At present most molecular datasets involve hundreds taxa whereas the current exact solution approaches have difficulty to tackle instances containing more than two dozen taxa (even smaller for the linear programming paradigm). Increasing the size of the datasets analyzable to the optimum is possibly one of the most challenging problems in molecular phylogenetics and warrants for sure further research efforts.

5 The likelihood paradigm of phylogenetic estimation

One of the most used criteria of phylogenetic estimation is the *likelihood criterion*. First formalized by Felsenstein [27], the likelihood criterion states that under many plausible explanations of an observed phenomenon, the one having the highest probability of occurring should be preferred to the others. When the likelihood criterion is applied to phylogenetic estimation, a phylogeny is defined to be optimal (or the most likely) if it has the highest probability of explaining the observed taxa. Thus, the likelihood paradigm consists of finding the phylogeny that maximizes a stochastic function, called *the likelihood function*, modeling a set of evolutionary hypotheses of the observed taxa.

The fundamental difference that distinguishes the likelihood paradigm from the least-squares and the minimum evolution paradigms is the nature of the information that it aims at finding. Specifically, if the least-squares and the minimum evolution paradigms aim at finding the best possible approximation of the projection of the evolutionary process of the observed taxa, the likelihood paradigm aims at reconstructing the most likely evolutionary

process that originated the observed taxa. Hence, if the phylogeny of the least-squares and the minimum evolution paradigms is an unrooted binary tree, the phylogeny of the likelihood paradigm is a rooted phylogeny, i.e., full binary tree having $(2n-1)$ vertices.

Formally, the likelihood function is defined to be a recursive function of a fixed rooted phylogeny T, a model of molecular evolution M, and an *observed data matrix* $\mathbf{S} = \{s_{rc}\}$, i.e., a matrix whose r-th row represents the molecular sequence of the r-th taxon. Defined the quantity

$$L_c^r(i) = \begin{cases} 1, \text{ if } s_{rc} = i \\ 0, \text{ otherwise,} \end{cases}$$

for each leaf r of T, each column c of $\mathbf{S}$, and each $i \in \Upsilon$, and the quantity

$$L_c^v(i) = \left[\sum_{j \in \Upsilon} L_c^{v_1}(j) p_{ij}(t_{v_1,v}) \right] \left[\sum_{j \in \Upsilon} L_c^{v_2}(j) p_{ij}(t_{v_2,v}) \right],$$

for each internal vertex v of T having v_1 and v_2 as children, the likelihood function $L(T, \mathbf{S}, M)$ of T can be defined as

$$L(T, \mathbf{S}, M) = \prod_c \left[\sum_{j \in \Upsilon} L_c^{\rho}(j) \pi_j \right],$$

where ρ denotes the root of T. In the context of the likelihood paradigm, the expected numbers of substitutions per site t_{v_h,v_k} assume the analogous meaning of edge weights in the least-squares and minimum evolution paradigms. Hence, when a given model of molecular evolution is assumed to hold (e.g., the THM model), finding the most likely phylogeny for a set of molecular sequences means maximizing the nonlinear (usually) non-convex stochastic function $L(T, \mathbf{S}, M)$ over all the possible rooted phylogenies, and for each rooted phylogeny, over all the possible associated edge weights t_{v_h,v_k} and substitution probabilities $p_{ij}(t_{v_h,v_k})$.

The NP-hardness of the likelihood paradigm [61] justified the development of a number of approximate solution approaches typically based on hill climbing strategies. Specifically, the strategies consist of a first phase in which the structure of a best-so-far phylogeny is modified and a second phase in which the nonlinear optimization of edge weights and the substitution probabilities is performed. The two phases are consecutively iterated until a stopping criterion is satisfied (e.g., the number of iterations performed or the elapsed time) [7, 27, 63]. A systematic review of the hill climbing strategies for the likelihood paradigm is out of the scope of the present chapter and can be found in Bryant et al. [7].

Recent mathematical advances on the likelihood paradigm led to overcome several limitations of the initial Felsenstein's model, such as the absence of

a rate variation among sites [80] and the absence of correlated evolution among sites [60]. Moreover, several progresses have been done concerning the analysis of its statistical consistency and its *idenfiability*, i.e., the study of the conditions under which the likelihood function is at least injective, an aspect markably related to its consistency [7]. The reader may find useful to refer to Gascuel [31] and Gascuel and Steel [33] for an overview these aspects.

5.1 The bayesian paradigm of phylogenetic estimation

Given a dataset of molecular sequences, suppose we have sufficient empirical evidence to assert that the evolution of the observed taxa followed a specific stochastic process. Then, we could try to combine this *a priori* information with the likelihood function in order to bias the search of the most probable phylogeny through those solutions that fit the known evolutionary process. This idea is at the core of the most recent likelihood-derived paradigm of phylogenetic estimation, called the *bayesian paradigm*, and will be briefly described in this section.

Similarly to the likelihood paradigm, the bayesian paradigm aims at finding the phylogeny that has the highest probability to recover the evolutionary process of the observed taxa. However, the selection of the most probable phylogeny is performed in light of the a priori information. Specifically, the a priori information is usually modeled by means of peculiar probability distributions, called *prior distributions*, that mainly concern three parameters, namely: the *topology*, i.e., the structure of the phylogeny, edge weights, and the substitution probabilities. Defined

$$\Theta = \{t_{v_h,v_k} \in \mathbb{R}_{0^+} : (v_k, v_k) \in T,\ \forall\ T \in \mathcal{T}\}$$

as the edge weight space and

$$\mathcal{R} = \{p_{ij}(t) \in [0,1] : \sum_{j \in \Upsilon} p_{ij}(t) = 1,\ \forall\ i,j \in \Upsilon,\ t \in \mathbb{R}_{0^+}\}$$

as the substitution probability space, the bayesian paradigm considers the prior distributions $\gamma(T)$, $\gamma(t)$, and $\gamma(R)$, to model the a priori information on $\mathcal{T}$, Θ, and $\mathcal{R}$, respectively. Selected an appropriate model of molecular evolution M, the prior distributions are then combined with the likelihood function to provide a *posterior density function* $B(T, \mathbf{S}, M)$ that represents the probability distribution of phylogenies conditional on the observed data matrix $\mathbf{S}$, the model M, and the priors distributions $\gamma(T)$, $\gamma(t)$, and $\gamma(R)$. Maximizing $B(T, \mathbf{S}, M)$ is the goal of the bayesian paradigm.

According to Bayes' theorem, fixed a phylogeny T_i and denoted t_i and R_i the corresponding subspaces of edge weights and substitution probabilities, the mathematical expression of the posterior probability $B(T_i, \mathbf{S}, M)$ of T_i

can be written as:

$$B(T_i, \mathbf{S}, M) = \frac{L_{\int}(T_i, \mathbf{S}, M)\gamma(T_i)}{\sum_{T_j \in \mathcal{T}} L_{\int}(T_j, \mathbf{S}, M)\gamma(T_j)}, \tag{17}$$

where $\gamma(T_i)$ denotes the prior probability of T_i, and $L_{\int}(T_i, \mathbf{S}, M)$ denotes the integral of the likelihood function $L(T_i, \mathbf{S}, M)$ over all possible edge weights and substitution probabilities [40], i.e.,

$$L_{\int}(T_i, \mathbf{S}, M) = \int_{t_i} \int_{R_i} L(T_i, \mathbf{S}, M)\gamma(t')\gamma(R')dt'dR'.$$

Hence, finding the optimal solution for the bayesian paradigm means finding the phylogeny T_i, the associated edge weights, and the substitution probabilities that globally maximize the posterior probability distribution of phylogenies $B(T, \mathbf{S}, M)$. Since finding the maximum a posteriori phylogeny implicitly implies being able to solve the likelihood paradigm, solving the bayesian paradigm is NP-hard [28].

The recursive nature of the likelihood function and the intractability of computing the denominator of Bayes' theorem prevent an analytical approach to solution of the bayesian paradigm. Hence, the maximum a posteriori phylogeny is usually computed by means of a *Markov chain Monte Carlo* (MCMC) algorithm [29], i.e., an algorithm that samples $B(T, \mathbf{S}, M)$ through a stochastic generation of phylogenies in $\mathcal{T}$ [see 48, 51, 82]. Sampling $B(T, \mathbf{S}, M)$ is extremely time consuming, therefore the bayesian estimations may take even weeks [41]. However, as observed by Yang [81] and Huelsenbeck et al. [40, 42], the sampling process has also the indisputable benefit of providing a measure of the reliability of the best-so-far solution found. In fact, by sampling stochastically around the (best local) maximum a posteriori phylogeny T^*, the bayesian paradigm could determine support values for the subtrees of T^*, i.e., measures of the posterior probability that the subtrees are true.

The bayesian paradigm is possibly the most complex among the phylogenetic estimation paradigms currently available in the literature on molecular phylogenetics. The recent computational advances obtained by Ronquist and Huelsenbeck [64] speeded up the execution of the MCMC algorithm and widened the use of the bayesian paradigm. However, the lack of a systematic investigation of its statistical consistency and the unclear dependence of the posterior density function on the a priori information [81] possibly make the bayesian paradigm still unripe for phylogenetic estimation [1].

5.2 Drawbacks of the likelihood and the bayesian paradigms of phylogenetic estimation

The higher the complexity of a paradigm, the higher the number of drawbacks that could arise, and the likelihood and the bayesian paradigms do not escape the rule. In fact, a number of computational and theoretical drawbacks affect the two paradigms. The computational drawbacks mainly involve the optimization aspects of the likelihood function and the sampling process in the bayesian paradigm. The theoretical drawbacks concern the evolutionary hypotheses at the core of the likelihood and bayesian criteria.

As regards to the computational drawbacks, in Section 5 we have seen that finding the most likely phylogeny for a set of taxa involves maximizing a nonlinear and generally non-convex stochastic function over all the possible phylogenies in $\mathcal{T}$, and for each phylogeny, over all the possible edge weights and substitution probabilities. Notoriously, this task can be only performed in an approximate way, due to a lack of general mathematical conditions that guarantee the global optimality of a solution in nonlinear non-convex programming [21, 53]. Hence, although it is possible (at least for small datasets) to enumerate all the possible phylogenies in $\mathcal{T}$, it is not possible to optimize globally edge weights and the substitution probabilities of a fixed phylogeny T. This fact may affect negatively the statistical consistency of the likelihood and the bayesian paradigms. In fact, the local optima of the likelihood function grows up exponentially in function of the number of taxa considered [7, 19, 20]. Thus, fixed a phylogeny T, the global optimum of the likelihood function is generally approximated by means of hill-climbing techniques that jump from local optimum to another one until a stopping criterion is satisfied (e.g., the number of iterations performed or the elapsed time) [7, 27, 63]. Assume that two phylogenies T_1 and T_2 are given, and let μ_1 and μ_2 be two vectors whose entries are edge weights and the substitution probabilities associated to T_1 and T_2, respectively. Let z_1 and z_2 the likelihood values of T_1 and T_2 for μ_1 and μ_2, respectively, and assume, without loss of generality, that $z_1 > z_2$. Due to the local nature of the optima μ_1 and μ_2, there could exists another local optimum, say $\hat{\mu}_2$, such that $\hat{z}_2 > z_1 > z_2$. If the hill-climbing algorithm finds $\hat{\mu}_2$ before μ_2 then we will consider T_2 as a better phylogeny than T_1, otherwise we will discard T_2 in favor of T_1. Hence, it is easy to realize that if one of the two phylogenies is the true phylogeny, its acceptance is subordinated to the goodness of the hill-climbing algorithm used to optimize the likelihood function, and as result the statistical consistency of the likelihood and bayesian paradigms may be seriously compromised.

Some authors argued that multiple local optima should arise infrequently in real datasets [63], but this conjecture was proved false by Bryant et al. [7] and Catanzaro et al. [12]. Specifically, Bryant et al. [7] observed that changing the model of molecular evolution influences the presence of multiple optima in the likelihood function, and Catanzaro et al. [12] showed a number of real

datasets affected by strong multimodality of the likelihood function. Despite the importance of the topic, to the best of our knowledge nobody was able to propose a plausible solution to this critical aspect.

A second computational drawback concerns the sampling process of the bayesian paradigm. In fact, as shown in Section 5.1, the approximation of the posterior density function is generally performed by means of a MCMC algorithm (e.g., the Metropolis or the Gibbs sampling algorithm [see 29]) that performs random walks in $\mathcal{T}$. The random walk should be sufficiently diversified to sample potentially the whole $\mathcal{T}$, and avoid double backs (i.e., to sample phylogenies already visited). Unfortunately, despite the recent computational advances in the bayesian paradigm [64], no technique may guarantee a sufficient diversification of the sampling process. Hence, the convergence to the maximum a posteriori phylogeny in practice becomes the convergence to the best-so-far a posteriori phylogeny that can be arbitrarily distinct from the true phylogeny [see 28, p. 296].

As regards to the theoretical drawbacks, it is worth noting that the evolutionary hypotheses at the core of the likelihood and bayesian criteria of phylogenetic estimation are at the same time their strength and their weakness. For example, if a proposed model of molecular evolution matches (at least roughly) the real evolutionary process of a set of molecular data then the likelihood and the bayesian paradigms could succeed in recovering the real phylogeny of the corresponding set of taxa (provided a solution to their computational drawbacks). However, if it is not the case, the paradigms will just provide a (sub)optimal solution for that model that may completely mismatch the real phylogeny. This aspect becomes evident e.g., in Rydin and Källersjö [66]'s article where, for a same dataset, two different Markov model of molecular evolution are used and two different maximum posterior phylogenies are obtained both having the 100% posterior probability of supporting the true phylogeny.

Finally, a second theoretical drawback concerns the prior distributions of the bayesian paradigm. In fact, it is worth noting that, if from one hand a strength of the bayesian paradigm is the ability to incorporate the a priori information, from the other hand this information is rarely available, hence in practical applications the prior distributions are generally modeled as uniform distributions, frustrating the potential strengths of the paradigm [1]. Moreover, it is unclear what type of information is well suited for a prior distribution; how possible conflicts among different sets of a priori information can be resolved; and if the inclusion of prior distributions strongly bias the estimation process. Huelsenbeck et al. [42] vaguely claimed "in a typical Bayesian analysis of phylogeny, the results are likely to be rather insensitive to the prior", but this results was not confirmed by Yang [81] who observed that "[...] the posterior probabilities of trees vary widely over simulated datasets [...] and can be unduly influenced by the prior [...]". Possibly, further research efforts are needed to provide answers to these practical concerns.

6 Conclusion

The success of a criterion of phylogenetic estimation is undoubtedly influenced by the quality of the evolutionary hypotheses at its core. If the hypotheses match (at least roughly) the real evolutionary process of a set of taxa, then the criterion will hopefully succeed in recovering the real phylogeny. Otherwise, the criterion will miserably fail, by suggesting an optimal phylogeny that mismatch partially or totally the correct result. Since we are far away from a complete understanding of the complex facets of evolution, it is not generally possible to assess the superiority of a criterion over others. Hence, families of estimation criteria cohabit in the literature of molecular phylogenetics, by providing different perspectives about the evolutionary process of the involved taxa.

In this chapter we have presented a general introduction of the existing literature about molecular phylogenetics. Our purpose has been to introduce a classification scheme in order to provide a general framework for papers appearing in this area. In particular, three main criteria of phylogenetic estimation have been outlined, the first based on the least-squares paradigm, firstly proposed by Cavalli-Sforza and Edwards [15], the second based on the minimum evolution paradigm, independently proposed by Kidd and Sgaramella-Zonta [44] and Beyer et al. [4], and the third based on the likelihood paradigm, firstly proposed by Felsenstein [27]. This division has been further disaggregated into different, approximately homogeneous subareas, and the basic aspects of each have been pointed out. For each, also, the most relevant issues affecting their use in tackling real-world sized problems have been outlined, as have the most interesting refinements deserving further research effort.

Acknowledgement

Daniele Catanzaro acknowledges support from the Belgian National Fund for Scientific Research (F.N.R.S.) of which he is “Chargé de Recherches”.

References

[1] J. K. Archibald, M. E. Mort, and D. J. Crawford. Bayesian inference of phylogeny: A non-technical primer. *Taxon*, 52:187–191, 2003.

[2] D. A. Bader, B. M. E. Moret, and L. Vawter. Industrial applications of high-performance computing for phylogeny reconstruction. In *SPIE ITCom: Commercial application for high-performance computing*, pages 159–168. SPIE, Bellingham, WA, 2001.

[3] J. P. Barthélemy and A. Guénoche. *Trees and proximity representations*. Wiley, New York, NY, 1991.

[4] W. A. Beyer, M. Stein, T. Smith, and S. Ulam. A molecular sequence metric and evolutionary trees. *Mathematical Biosciences*, 19:9–25, 1974.

[5] Å. Björck. *Numerical Methods for Least-Squares Problems*. SIAM, Philadelphia, PA, 1996.

[6] J. Brinkhuis and V. Tikhomirov. *Optimization: Insights and applications*. Princeton University Press, Princeton, NJ, 2005.

[7] D. Bryant, N. Galtier, and M. A. Poursat. Likelihood calculation in molecular phylogenetics. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*. Oxford University Press, New York, NY, 2005.

[8] R. M. Bush, C. A. Bender, K. Subbarao, N. J. Cox, and W. M. Fitch. Predicting the evolution of human influenza A. *Science*, 286(5446):1921–1925, 1999.

[9] D. Catanzaro. The minimum evolution problem: Overview and classification. *Networks*, 53(2):112–125, 2009.

[10] D. Catanzaro, L. Gatto, and M. Milinkovitch. Assessing the applicability of the GTR nucleotide substitution model through simulations. *Evolutionary Bioinformatics*, 2, 2006.

[11] D. Catanzaro, R. Pesenti, and M. Milinkovitch. A non-linear optimization procedure to estimate distances and instantaneous substitution rate matrices under the GTR model. *Bioinformatics*, 22(6):708–715, 2006.

[12] D. Catanzaro, R. Pesenti, and M. C. Milinkovitch. A very large-scale neighborhood search to estimate phylogenies under the maximum likelihood criterion. Technical report, G.O.M. - Computer Science Department - Université Libre de Bruxelles (U.L.B.), 2007.

[13] D. Catanzaro, M. Labbé, R. Pesenti, and J. J. Salazar-Gonzalez. The balanced minimum evolution problem. Technical report, G.O.M. - Computer Science Department - Université Libre de Bruxelles (U.L.B.), 2009.

[14] D. Catanzaro, M. Labbé, R. Pesenti, and J. J. Salazar-Gonzalez. Mathematical models to reconstruct phylogenetic trees under the minimum evolution criterion. *Networks*, 53(2):126–140, 2009.

[15] L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic analysis: Models and estimation procedures. *American Journal of Human Genetics*, 19:233–257, 1967.

[16] R. Chakraborty. Estimation of time of divergence from phylogenetic studies. *Canadian Journal of Genetics and Cytology*, 19:217–223, 1977.

[17] B. S. W. Chang and M. J. Donoghue. Recreating ancestral proteins. *Trends in Ecology and Evolution*, 15(3):109–114, 2000.

[18] L. Chisci. *Sistemi Dinamici - Parte I*. Pitagora, Bologna, Italy, 2001.

[19] B. Chor, M. D. Hendy, B. R. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Molecular Biology and Evolution*, 17(10):1529–1541, 2000.

[20] B. Chor, M. D. Hendy, and S. Snir. Maximum likelihood jukes-cantor triplets: Analytic solutions. *Molecular Biology and Evolution*, 23(3):626–632, 2005.

[21] A. R. Conn, N. I. M. Gould, and P. L. Toint. *Trust-region methods*. SIAM, Philadelphia, PA, 2000.

[22] W. H. E. Day. Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49:461–467, 1987.

[23] F. Denis and O. Gascuel. On the consistency of the minimum evolution principle of phylogenetic inference. *Discrete Applied Mathematics*, 127: 66–77, 2003.

[24] R. Desper and O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. *Journal of Computational Biology*, 9(5):687–705, 2002.

[25] R. Desper and O. Gascuel. Theoretical foundations of the balanced minimum evolution method of phylogenetic inference and its relationship to the weighted least-squares tree fitting. *Molecular Biology and Evolution*, 21(3):587–598, 2004.

[26] M. Farach, S. Kannan, and T. Warnow. A robust model for finding optimal evolutionary trees. *Algorithmica*, 13:155–179, 1995.

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

[28] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, Sunderland, MA, 2004.

[29] G. S. Fishman. *Monte Carlo: Concepts, algorithms, and applications*. Springer-Verlag, New York, NY, 1996.

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

[31] O. Gascuel. *Mathematics of evolution and phylogeny*. Oxford University Press, New York, NY, 2005.

[32] O. Gascuel and D. Levy. A reduction algorithm for approximating a (non-metric) dissimilarity by a tree distance. *Journal of Classification*, 13:129–155, 1996.

[33] O. Gascuel and M. A. Steel. *Reconstructing evolution*. Oxford University Press, New York, NY, 2007.

[34] O. Gascuel, D. Bryant, and F. Denis. Strengths and limitations of the minimum evolution principle. *Systematic Biology*, 50:621–627, 2001.

[35] P. H. Harvey, A. J. L. Brown, J. M. Smith, and S. Nee. *New uses for new phylogenies*. Oxford University Press, Oxford, UK, 1996.

[36] M. Hasegawa, H. Kishino, and T. Yano. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

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

[38] D. P. Heyman and M. J. Sobel, editors. *Stochastic models*, volume 2 of *Handbooks in operations research and management science*. North-Holland, Amsterdam, The Netherlands, 1990.

[39] S. Horai, Y. Sattah, K. Hayasaka, R. Kondo, T. Inoue, T. Ishida, S. Hayashi, and N. Takahata. Man's place in the hominoidea revealed by mitochondrial dna genealogy. *Journal of Molecular Evolution*, 35:32–43, 1992.

[40] J. P. Huelsenbeck, B. Larget, P. van der Mark, and F. Ronquist. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8): 754–755, 2001.

[41] 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.

[42] J. P. Huelsenbeck, B. Larget, R. E. Miller, and F. Ronquist. Potential applications and pitfalls of bayesian inference of phylogeny. *Systematic Biology*, 51:673–688, 2002.

[43] T. H. Jukes and C.R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian Protein Metabolism*, pages 21–123. Academic Press, New York, NY, 1969.

[44] K. K. Kidd and L. A. Sgaramella-Zonta. Phylogenetic analysis: Concepts and methods. *American Journal of Human Genetics*, 23:235–252, 1971.

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

[46] M. K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal rates. *Molecular Biology and Evolution*, 11(3):584–593, 1994.

[47] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93, 1984.

[48] S. Li, D. Pearl, and H. Doss. Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association*, 95:493–508, 2000.

[49] V. Makarenkov and B. Leclerc. An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. *Journal of Classification*, 16:3–26, 1999.

[50] M. A. Marra, S. J. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, A. Cloutier, S. M. Coughlin, D. Freeman, N. Girn, O. L. Griffith, S. R. Leach, M. Mayo, H. McDonald, S. B. Montgomery, P. K. Pandoh, A. S. Petrescu, A. G. Robertson, J. E. Schein, A. Siddiqui, D. E. Smailus, J. M. Stott, G. S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T. F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G. A. Tipples,

S. Tyler, R. Vogrig, D. Ward, B. Watson, R. C. Brunham, M. Krajden, M. Petric, D. M. Skowronski, C. Upton, and R. L. Roper. The genome sequence of the SARS-associated coronavirus. *Science*, 300(5624):1399–1404, 2003.

[51] B. Mau and M. A. Newton. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 6:122–131, 1997.

[52] G. L. Nemhauser and L. A. Wolsey. *Integer and combinatorial optimization*. Wiley-Interscience, New York, NY, 1999.

[53] G. L. Nemhauser, A. H. G. Rinnooy Kan, and M. J. Tod, editors. *Optimization*, volume 1 of *Handbooks in operations research and management science*. North-Holland, Amsterdam, The Netherlands, 1989.

[54] C. Y. Ou, C. A. Ciesielski, G. Myers, C. I. Bandea, C. C. Luo, B. T. M. Korber, J. I. Mullins, G. Schochetman, R. L. Berkelman, A. N. Economou, J. J. Witte, L. J. Furman, G. A. Satten, K. A. Maclnnes, J. W. Curran, and H. W. Jaffe. Molecular epidemiology of HIV transmission in a dental practice. *Science*, 256(5060):1165–1171, 1992.

[55] L. Pachter and B. Sturmfels. The mathematics of phylogenomics. *SIAM Review*, 49(1):3–31, 2007.

[56] R. D. M. Page and E. C. Holmes. *Molecular Evolution: A Phylogenetic Approach*. Blackwell Science, Oxford, UK, 1998.

[57] J. M. Park and M. W. Deem. Phase diagrams of quasispecies theory with recombination and horizontal gene transfer. *Physical Review Letters*, 98: 058101–058104, 2007.

[58] Y. Pauplin. Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 51:41–47, 2000.

[59] P. A. Pevzner. *Computational molecular biology*. MIT Press, Cambridge, MA, 2000.

[60] D. D. Pollock, W. R. Taylor, and N. Goldman. Coevolving protein residues: Maximum likelihood identification and relationship to structure. *Journal of Molecular Biology*, 287(1):187–198, 1999.

[61] S. Roch. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(1):92–94, 2006.

[62] F. Rodriguez, J. L. Oliver, A. Marin, and J. R. Medina. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142:485–501, 1990.

[63] J. S. Rogers and D. Swofford. Multiple local maxima for likelihoods of phylogenetic trees from nucleotide sequences. *Molecular Biology and Evolution*, 16:1079–1085, 1999.

[64] F. Ronquist and J. P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.

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

[66] C. Rydin and M. Källersjö. Taxon sampling and seed plant phylogeny. *Cladistics*, 18:485–513, 2002.

[67] A. Rzhetsky and M. Nei. Theoretical foundations of the minimum evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10:1073–1095, 1993.

[68] A. Rzhetsky and M. Nei. Statistical properties of the ordinary least-squares generalized least-squares and minimum evolution methods of phylogenetic inference. *Journal of Molecular Evolution*, 35:367–375, 1992.

[69] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4: 406–425, 1987.

[70] E. Schadt and K. Lange. Codon and rate variation models in molecular phylogeny. *Molecular Biology and Evolution*, 19(9):1534–1549, 2002.

[71] E. Schadt and K. Lange. Applications of codon and rate variation models in molecular phylogeny. *Molecular Biology and Evolution*, 19(9):1550–1562, 2002.

[72] C. Semple and M. A. Steel. *Phylogenetics*. Oxford University Press, New York, NY, 2003.

[73] P. H. A. Sneath and R. R. Sokal. *Numerical taxonomy*. W. K. Freeman and Company, San Francisco, CA, USA, 1963.

[74] J. A. Studier and K. J. Keppler. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5:729–731, 1988.

[75] D. L. Swofford. *PAUP* version 4.0*. Sinauer Associates, Sunderland, MA, 1997.

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

[77] 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(3):512–526, 1993.

[78] P. J. Waddell and M. A. Steel. General time-reversible distances with unequal rates across sites: Mixing gamma and inverse gaussian distributions with invariant sites. *Molecular Phylogenetics and Evolution*, 8: 398–414, 1997.

[79] M. S. Waterman, T. F. Smith, M. Singh, and W. A. Beyer. Additive evolutionary trees. *Journal of Theoretical Biology*, 64:199–213, 1977.

[80] Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39:306–314, 1994.

[81] Z. Yang. Bayesian inference in molecular phylogenetics. In O. Gascuel, editor, *Mathematics of Evolution and Phylogeny*. Oxford University Press, New York, NY, 2005.

[82] Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Molecular Biology and Evolution*, 14:717–724, 1997.

[83] L. A. Zadeh and C. A. Desoer. *Linear system theory*. McGraw-Hill, New York, NY, 1963.