Distance-based methods in phylogenetics

Fabio Pardi, Olivier Gascuel

Distance-based methods in phylogenetics

Fabio Pardi* and Olivier Gascuel*

*: Institut de Biologie Computationnelle, LIRMM UMR 5506, CNRS & Université de Montpellier, France

Fabio Pardi
Email: pardi@lirmm.fr
Phone: +33 (0)4 6714 9776
Address: LIRMM, Campus St Priest - BAT 5
860 rue de St Priest
34095 Montpellier cedex 5
France

Olivier Gascuel
Email: gascuel@lirmm.fr
Phone: +33 (0)4 6741 8547
Address: LIRMM, Campus St Priest - BAT 5
860 rue de St Priest
34095 Montpellier cedex 5
France

Glossary

Additive distances: a collection of pairwise distances δ_{ij} is additive if there exists a phylogenetic tree whose branch lengths determine distances between leaves coinciding with the δ_{ij}.

Branch lengths: the branches in a phylogenetic tree usually have lengths representing the amount of evolutionary change that has occurred between the taxa at their endpoints. When the tree represents the evolution of a set of biological sequences, branch lengths are usually measured in terms of expected number of substitutions per site.

Computational complexity: a measure of how a resource (typically time or memory) employed by an algorithm scales with the input data. An algorithm has complexity $O(f(n))$ if the resources grow at most proportionally to $f(n)$, a function of the input size n.

Distance matrix: a square matrix, i.e., a table with an equal number of rows and columns corresponding to a set of taxa, where the entry in column i and row j, denoted δ_{ij} represents an estimate of the evolutionary distance between the i-th and the j-th taxon. The matrix is symmetric (i.e., $\delta_{ij} = \delta_{ji}$) and the elements on its diagonal are zero ($\delta_{ii} = 0$).

Evolutionary distance: a measure of the amount of change or divergence that has occurred between two taxa. When taxa are represented by biological sequences, `change' is usually taken as the occurrence of substitutions along the evolutionary path connecting the sequences.

Least squares (LS): a general principle from regression analysis to adjust the parameters of a model function to best fit some observed data. In distance-based reconstruction, it is used to fit the branch lengths of a phylogeny to a set of distance estimates. Least squares methods aim to minimize a quadratic function of the differences between the distance estimates and the additive distances determined by the branch lengths. For *ordinary least squares* (OLS) the goal is to minimize the sum of the squared differences, while for *weighted least squares* (WLS) this sum is weighted by terms reflecting the variances of the distance estimates.

Minimum evolution (ME): a general distance-based principle, analogous to parsimony, to measure the plausibility of a phylogenetic tree whose branch lengths have been fitted to the data, typically by least squares (see above). If we define the length of a tree as the sum of its branch lengths, then shorter trees are deemed to be more plausible than longer ones.

Molecular clock: the assumption of a constant substitution rate across a phylogeny, implying that branch lengths are directly proportional to time.

Molecular phylogenetics: the study and reconstruction of phylogenies representing the evolution of molecular sequences, such as those of DNA or proteins.

Nearest neighbor interchange (*NNI*): a rearrangement of a tree topology swapping the position of two subtrees separated by exactly 3 branches. For example, one NNI can transform tree ((A,B),(C,D)) into ((A,C),(B,D)) or into ((A,D),(B,C)).

Sequence profile: a table describing the general form of a collection of aligned sequences. For each site, the profile specifies the frequency of each possible character (including gaps) at that position.

Statistical consistency: in statistics, a method of estimation is statistically consistent if it is guaranteed to converge, with probability 1, towards the correct value of the parameter, as the size of the input sample tends to infinity. In molecular phylogenetics, the central parameter is often taken to be the topology τ of the correct phylogenetic tree, so a method is said to be statistically consistent if the probability of reconstructing a tree of topology τ converges to 1 as the input sequence alignment becomes longer and longer.

Substitution model: a probabilistic model describing the occurrence of substitutions in a DNA or protein sequence. Substitutions models are used to estimate the evolutionary distances between pairs of molecular sequences.

Subtree pruning and regrafting (*SPR*): a rearrangement of a tree topology, which removes a subtree (a clade) and reinserts it elsewhere in the tree topology.

Tree topology: the discrete, structural information contained in a phylogenetic tree, besides branch length information.

Ultrametric distances: the pairwise distances δ_{ij} are ultrametric if they are additive with respect to a tree whose leaves are all at the same distance from the root. In molecular phylogenetics, distances are approximately ultrametric only when the molecular clock assumption holds.

Keywords

Additive distance, branch lengths, computational complexity, consistency, distance matrix, evolutionary distance, least squares, minimum evolution, neighbor joining, phylogenetic tree.

Synopsis

A popular approach in phylogenetics consists in estimating a matrix of evolutionary distances between pairs of taxa, and then using this information to reconstruct a phylogenetic tree for those taxa. In this entry, we first explain how distances should be defined and estimated, and then focus on the task of inferring a phylogenetic tree that accounts for the estimated distances. Specifically, we will introduce the classical guiding principles for tree inference – least squares and minimum evolution – and then present the most popular distance-based methods – in particular neighbor joining and a wide array of algorithms inspired by it.

1. Introduction

Distance-based phylogenetic reconstruction rests on two steps. First, for each pair of taxa estimate the amount of change that has occurred along the evolutionary path connecting them. Such measure of change is called an *evolutionary distance*, and in the case of biological (DNA or protein) sequences, it is proportional to the number of substitutions that have taken place in the two lineages since the last common ancestor of the two sequences. Tree inference is then based on these estimates – the goal being to find the tree that best accounts for the estimated distances.

Using pairwise distances or similarities between taxa is arguably the oldest approach to investigate systematic relationships by means of a computer (Sneath 1957, Sokal & Michener 1958). Early work saw a very clear distinction between, on the one hand, hierarchical clustering techniques for taxonomic classification (Sokal & Sneath 1963), and, on the other hand, optimization-based methods directly aimed at phylogenetic reconstruction (Cavalli-Sforza & Edwards 1967, Fitch & Margoliash 1967). These two lines of work eventually converged in the 1980s, leading to the hugely popular neighbor joining (`NJ`) method (Saitou & Nei 1987) – which combined algorithmic ideas from classification (e.g., `UPGMA`, Sokal & Michener 1958; `ADDTREE`, Sattath & Tversky 1977) to optimization principles from phylogenetics (e.g. ME, Kidd & Sgaramella-Zonta 1971; OLS, Cavalli-Sforza & Edwards 1967; see below). From `NJ` onwards, phylogenetics has witnessed a renaissance of distance-based methods, often related to, or inspired by `NJ` (e.g., `FastME`, Desper & Gascuel 2002; `FastNJ`, Elias & Lagergren 2009; `FastTree`, Price et al. 2009).

These methods are still widely used for their computational efficiency, an advantage that makes them particularly suited for the reconstruction of very large phylogenies, or large collections of phylogenies (e.g., for bootstrapping), or to provide a basis for progressive multiple sequence alignments (Larkin et al. 2007), or even to construct initial trees for more sophisticated inference approaches such as those based on maximum likelihood (Guindon & Gascuel 2003).

More generally, using estimated distances between biological sequences is an obvious answer to cope with the massive data sets generated by the modern, ever faster and cheaper sequencing techniques – as testified by the ongoing success of NJ, which to date remains the most cited algorithm in phylogenetics.

In this entry, we outline the main ideas that underlie the methodology for distance-based phylogenetics. We start by explaining the importance of estimating distances that reflect the number of changes that have actually occurred between two taxa, rather than the (smaller) number of differences between them. After describing concisely the task of estimating evolutionary distances, the main focus here is on the methods for tree inference proper, that is, to reconstruct a tree that fits well the estimated distances.

2. Preliminaries

A *phylogenetic tree* T over a set of taxa X has two components. First, a tree topology τ, i.e. an unrooted tree with no degree-2 nodes, whose leaves are labeled by (and represent) the taxa in X, and whose internal nodes represent putative ancestors of these taxa. The second component consists of positive branch lengths $b(e)$ for every branch e in τ, which represent a measure of evolutionary change occurred along e.

Every phylogenetic tree T determines a collection of *tree distances* d_{ij}^T between the taxa $i, j \in X = \{1,2, \dots, n\}$ labeling its leaves. They are defined by:

$$d_{ij}^T = \sum_{e \in \tau_{ij}} b(e),$$

where τ_{ij} denotes the set of branches between i and j in τ. In other words, the tree distance d_{ij}^T is the length of the path connecting i and j in T.

Now suppose that, for each pair of taxa $i, j \in X = \{1,2, \dots, n\}$, an estimate δ_{ij} of their evolutionary distance is obtained. We say that the δ_{ij} are *additive*, if $\delta_{ij} = d_{ij}^T$ for all $i, j \in X$, for some phylogenetic tree T over X. Importantly, when such a tree exists, it is unique (Zaretskii 1965, Simões Pereira 1969, Buneman 1971). Moreover, as we shall see in the following, it is algorithmically easy to reconstruct tree T from its tree distances d_{ij}^T. These observations provide the fundamental idea underlying distance-based methods: if we manage to obtain precise estimates of the distances d_{ij}^T for the phylogenetic tree we seek, then reconstructing this tree is easy.

The word additive comes from the fact that if we could observe a taxon i as it evolves into k at an intermediate stage, and then eventually into j, then the true evolutionary distances between i, j, k must satisfy

$$d_{ij} = d_{ik} + d_{kj}. \qquad (1)$$

```
i:  GAATACTCAAA
    ||.|.|.|.||
k:  GACTGCCCGAA
    ||.|||.||.|
j:  GATTGCTCGGA
```

Figure 1: **Uncorrected distances are not additive.** We assume that i, k and j are realizations of the same sequence at three successive times. If distances were defined as the number of differences, we would have $d_{ij} = 4 < d_{ik} + d_{kj} = 7$, contradicting equation (1), and showing that the distances would not be additive. Note that both d_{ij} and $d_{ik} + d_{kj}$ potentially underestimate the number of substitutions occurred between i and j.

3. Distance estimation

The first, fundamental component of a distance-based method is the definition of `evolutionary distance'. It is important to understand that not any measure of distance can be adopted for phylogenetic reconstruction: the key requirement is that the parameters that we set out to estimate must be additive in the sense specified above. This ensures that, as the data become more and more abundant and the distance estimates more and more accurate, these estimates will determine the correct phylogenetic tree (Atteson 1999).

As a consequence, in molecular phylogenetics, where the data are collections of DNA or protein sequences, simply counting the number of differences between each pair of sequences is not acceptable, because of the possibility of multiple substitutions at the same site (see Fig. 1). The number of differences, or mismatches, between sequences is sometimes referred to as their `uncorrected distance'.

A much better approach is to define the distances as (proportional to) the number of substitutions that have occurred between the two sequences, which clearly leads to an additive measure. As this number is unobservable, the general approach is to estimate it using nucleotide or amino acid substitution models. Note that uncorrected distances do not account for unobserved changes – such as multiple substitutions at the same site – and therefore underestimate this number (see again Fig. 1). We will now describe, in very general terms, the maximum likelihood (ML) approach to solve this estimation problem. The interested reader is referred to more advanced textbooks (e.g. Felsenstein 2004, Yang 2006) for a detailed treatment of distance estimation.

Substitution models allow us to calculate a substitution probability matrix $P(d) = (p_{xy}(d))$, where x and y denote nucleotides, amino acids or other biological characters, and $p_{xy}(d)$ denotes the probability that an x becomes a y after evolving along a branch of length d. Note that d is not expressed in units of time. Instead, the rate of substitution models is usually scaled so that d equals the expected number of substitutions per site along a branch of that length. Also recall that π_x denotes the stationary probability of x. It can be defined as $\pi_x = \lim_{d\to\infty} p_{yx}(d)$, and is sometimes estimated using the frequency of x in the sequences being analyzed.

The evolutionary distance between two sequences x and y is estimated on the basis of a pairwise alignment of these two sequences. Denote by x_i and y_i the i-th aligned character of x and y, respectively. Assuming that the substitution model is time-reversible, the likelihood is given by:

$$L(d) = \prod_{i=1}^{m} \pi_{x_i} p_{x_i y_i}(d),$$

where the product is over the m aligned sites. Then, the ML estimate of the distance between x and y is the value of d that maximizes $L(d)$ above, and can be obtained numerically or analytically, depending on the model.

For illustration, we consider the simplest model of nucleotide substitution, the JC model (Jukes and Cantor 1969). For this model, we have, $\pi_A = \pi_C = \pi_G = \pi_T = {}^1\!/_4$ and, assuming $x \neq y$,

$$p_{xx}(d) = \frac{1}{4}\left(1 + 3e^{-\frac{4}{3}d}\right), \qquad p_{xy}(d) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}d}\right).$$

The likelihood is then given by:

$$L(d) = \frac{1}{4^{2m}}\left(1 - e^{-\frac{4}{3}d}\right)^{m_{\neq}} \left(1 + 3e^{-\frac{4}{3}d}\right)^{m-m_{\neq}},$$

where $m_{\neq}$ is the number of mismatches. It is then easy to calculate that $L(\delta)$ is maximized for

$$\delta = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{m_{\neq}}{m}\right).$$

That is, the distance estimate δ between the two sequences is a simple, strictly increasing function of the proportion of mismatches $m_{\neq}/m$ – the uncorrected distance we mentioned above. Note that $\delta \geq m_{\neq}/m$, corresponding to the fact that there are more substitutions than observed differences.

Other models cause the ML estimates of the distances to be functions of multiple features of the pairwise alignment, so strictly speaking it is not always accurate to describe distance estimates as transformations of the uncorrected distances. As ML distance estimation is the same as ML phylogenetic reconstruction of a 2-taxon tree, the numerical techniques for distance estimation are largely the same as those employed for ML branch length optimization (Yang 2006).

4. Tree reconstruction

We organize our brief survey of tree reconstruction methods around three well-defined components. Any choice with respect to them defines a possible distance-based method.

C1 *Branch length estimation*: a method to assign lengths to the branches of any fixed tree topology, so that the resulting tree distances are as close as possible to the estimated distances δ_{ij}. This is usually achieved using least squares techniques from regression analysis (Cavalli-Sforza & Edwards 1967, Fitch & Margoliash 1967).

C2 *What to optimize*: a criterion assigning a score to all the trees of different topologies that can be obtained with component C1, reflecting the biological plausibility of a phylogenetic reconstruction given the estimated distances. An obvious choice for this is the least squares criterion used to assign branch lengths, but as we describe below a lot of recent methodology is based on a different criterion, *minimum evolution* (Kidd & Sgaramella-Zonta 1971). Some methods (e.g. ADDTREE; Sattath & Tversky 1977) directly optimize topological criteria, and thus bypass C1.

C3 *How to optimize*: an algorithm to seek the optimal tree with respect to the criterion in C2. Since this is a computationally hard optimization problem, algorithms are usually heuristic, and based on simple but effective ideas such as stepwise addition, iterative agglomeration, or hill climbing (Swofford et al. 1990, Felsenstein 2004).

In the following, we start by surveying the methodology for C1 (Sec. 4.1). Then we describe one of the most popular criteria for C2, minimum evolution (Sec. 4.2), which is at the foundation of what is still the best known distance-based method, neighbor joining. We illustrate this algorithm along with other approaches that are based on the same ideas (Sec. 4.3). Finally, we briefly describe a few promising approaches which, strictly speaking, are not distance-based methods, but which share with them several ideas and the same emphasis on computational efficiency (Sec. 4.4).

4.1 Least squares branch length estimation

Given the estimated distances δ_{ij} for all $i, j \in X$, the goal of least squares phylogenetic reconstruction (Cavalli-Sforza & Edwards 1967, Fitch & Margoliash 1967) is to find a phylogenetic tree T that minimizes the gap between the estimates δ_{ij} and the tree distances d_{ij}^T, measured in terms of a quadratic function $Q(T)$ of the residuals $\delta_{ij} - d_{ij}^T$. Different, statistically motivated choices for $Q(T)$ are possible, and are detailed below. While many versions of this problem have been proven computationally hard (Day 1987), here we focus on the simpler problem of assigning branch lengths to a tree of fixed topology τ. As we show below, exact, analytic, and polynomially-computable solutions are available for this task.

The method of least squares was first introduced in phylogenetics in the mid 1960s. The proposed objective function was

$$Q(T) = \sum_{i<j} w_{ij}\left(\delta_{ij} - d_{ij}^T\right)^2,$$

with $w_{ij} = 1$ (Cavalli-Sforza & Edwards 1967) and $w_{ij} = 1/\delta_{ij}^2$ (Fitch & Margoliash 1967). The former is often referred to as *ordinary least squares* (OLS), and both approaches are particular cases of *weighted least squares* (WLS), where various choices are possible for w_{ij}. From a statistical standpoint, the weight w_{ij} represents the degree of confidence that we can attach to the distance estimate δ_{ij}: ideally w_{ij} should be inversely proportional to the variance of δ_{ij}, but in practice setting the weights is a delicate art, because the variances are difficult to evaluate. One particular choice on which we will come back a few times in the following is that of *balanced* weights – with w_{ij} proportional to $2^{-|\tau_{ij}|}$ ($|\tau_{ij}|$ denotes the number of edges on the path between i and j in τ) – which assigns less confidence to the distances between topologically distant taxa.

WLS approaches ignore the correlations between the distance estimates δ_{ij}, δ_{kl} for different pairs of taxa, which may be significant when the paths τ_{ij} and τ_{kl} share many branches. In order to take these correlations into account, *generalized least squares* (GLS) (Chakraborty 1977, Bulmer 1991) minimizes

$$Q(T) = \sum_{i<j}\sum_{k<l} w_{ij,kl}\left(\delta_{ij} - d_{ij}^T\right)\left(\delta_{kl} - d_{kl}^T\right),$$

where the $w_{ij,kl}$ should be set as the entries of the inverse of the variance-covariance matrix for the distance estimates δ_{ij}. Just as OLS is a particular case of WLS, WLS is a particular case of GLS, obtained by setting $w_{ij,kl} = 0$ whenever $\{i,j\} \neq \{k,l\}$. In practice GLS is rarely used for phylogenetic inference, because of the difficulty of evaluating the covariances, and because of its higher computational costs relative to those of OLS and WLS.

The branch lengths that are optimal with respect to the $Q(T)$ criteria above can be simply expressed in matrix notation. To this end, let $\boldsymbol{\delta} = (\delta_{ij})$ denote the distance estimates and $\boldsymbol{d^T} = (d_{ij}^T)$ the tree distances for a tree T, in vector form. Moreover, we represent any topology τ with a binary matrix $A_\tau = (a_{ij,e})$ – whose rows correspond to pairs of taxa $\{i,j\}$ and whose columns correspond to branches of τ – defined by setting $a_{ij,e} = 1$ if e is on the path between i and j in τ, and 0 otherwise. Given these notations, we can write

$$\boldsymbol{d^T} = A_\tau \boldsymbol{b},$$

where $\boldsymbol{b} = (b(e))$ denotes the branch lengths of T in vector form. See Fig. 2 for an example illustrating these notations.

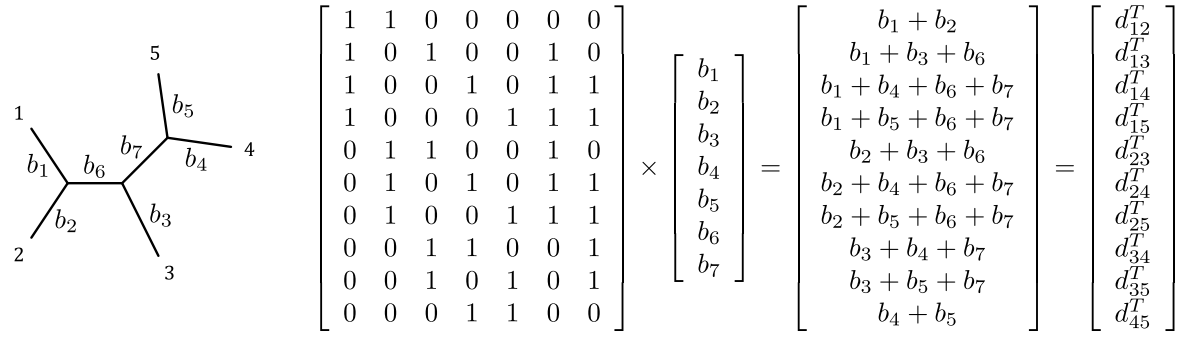

Figure 2: **The usefulness of the topological matrix.** The tree distances for the 5-taxon tree on the left are expressed here as $A_\tau \boldsymbol{b} = \boldsymbol{d^T}$.

The objective functions of OLS, WLS and GLS can then be written concisely in matrix form:

$$Q(T) = (\boldsymbol{\delta} - A_\tau \boldsymbol{b})^t W (\boldsymbol{\delta} - A_\tau \boldsymbol{b}),$$

where $W = (w_{ij,kl})$ contains the GLS weights and the superscript t denotes the matrix transpose. When W is diagonal, or the identity matrix, WLS and OLS are obtained, respectively. The branch lengths for τ that minimize $Q(T)$ can then be expressed as:

$$\boldsymbol{b} = (A_\tau^t W A_\tau)^{-1} A_\tau^t W \boldsymbol{\delta}. \qquad (2)$$

The matrix calculations in equation (2) are computationally expensive, but a number of properties of the matrices involved can be exploited to speed up the solution (Gascuel 1997a, Bryant & Waddell 1998). For WLS, computational complexity is dominated by the matrix inversion – or equivalently by linear system resolution – which standard algorithms achieve in $O(n^3)$ time, where we recall that $n = |X|$ denotes the number of taxa. For OLS, the complexity can be reduced further to $O(n^2)$. In fact, the OLS branch lengths can be expressed with simple combinatorial formulae (Vach 1989, Rzhetsky & Nei 1993). More recently it was discovered that similar formulae exist for WLS with balanced weights (Desper & Gascuel 2004), and in fact for a whole new class of special cases of WLS, those with *multiplicative* weights (Mihaescu & Pachter 2008).

We conclude by remarking that none of the approaches described above guarantees that the assigned branch lengths are all positive. Least squares branch length estimates are sometimes negative, which does not correspond to any biological reality. Constraining branch lengths to be non-negative leads to non-negative least-squares (NNLS) regression (Lawson & Hanson 1974), an approach that however further increases computational costs.

4.2 Optimization criteria: minimum evolution

Once a way to assign branch lengths to any fixed topology has been determined, the next question (C2) is how to measure the plausibility of the trees we obtain. One possibility is to adopt the same least squares criterion $Q(T)$ used for branch lengths, implying that a tree topology is considered plausible if its tree distances can be fitted very closely to the estimated distances. This approach is known to perform best when negative branch lengths are disallowed (Kuhner & Felsenstein 1994), a constraint that can be imposed in popular programs such as FITCH (Felsenstein 1997) and PAUP* (Swofford 2003).

A different approach – *minimum evolution* (ME) – scores a tree using the sum $L(T)$ of the fitted branch lengths: the shorter the tree, the better. The intuition underlying ME is the same as that of maximum parsimony for character-based tree reconstruction: similarly to the general principle that simple explanations are preferable to complex ones, in phylogenetics shorter trees are often considered more plausible than longer ones.

ME can deal with negative branch lengths in a variety of ways: they can be simply excluded from the sum in $L(T)$ (Swofford et al. 1990), or $L(T)$ can be defined as the sum of the absolute values of the branch lengths (Kidd & Sgaramella-Zonta 1971). To date, the most common approach is to define $L(T)$ as the sum of all branch lengths, irrespective of their sign (Saitou & Imanishi 1989, Rzhetsky & Nei 1993), which would seem to favor negative branch lengths, but in practice works quite well.

The first theoretical foundation of ME was provided by Rzhetsky and Nei, who showed that, if the distance estimates are unbiased, and branch lengths are assigned with OLS, then the mathematical expectation of $L(T)$ is minimized for the correct tree topology (Rzhetsky & Nei 1993). This means that, when the distance estimates equal the tree distances for a tree T – that is when $\boldsymbol{\delta} = \boldsymbol{d^T}$ – then the optimal tree with respect to OLS+ME is T itself. Note that in the

following we will use the short form `C1+C2' to denote the optimization principle based on a particular choice for components C1 and C2 (see above).

If we assume that $\boldsymbol{\delta}$ converges to $\boldsymbol{d^T}$ as more and more data are available, the result above implies that OLS+ME is *statistically consistent* – meaning that the probability of reconstructing T (within any given approximation) converges to 1 – which is an essential property of any phylogenetic inference method. Unfortunately ME does not always have this property (Gascuel et al. 2001) – an observation that casts serious doubts on the general applicability of the ME principle in phylogenetics. To date, statistical consistency has been proven to hold for all instances of WLS+ME with multiplicative weights (Pardi & Gascuel 2012).

One special case of this is WLS+ME with balanced weights, which is also known as BME (Pauplin 2000, Desper & Gascuel 2002, 2004). This optimization principle has several interesting mathematical features (Semple and Steel 2004), including the fact that for fully resolved trees its objective function can be expressed very concisely – and elegantly – as a function of the estimated distances:

$$L(T) = \sum_{i<j} 2^{1-|\tau_{ij}|}\, \delta_{ij} \tag{3}$$

BME is of key importance to interpret some of the most central methods in distance-based phylogenetics, including neighbor joining, as we explain below.

4.3 NJ and related algorithms

Neighbor joining (NJ) is an agglomerative clustering algorithm, that is, it constructs a tree in a bottom-up fashion by alternating the following two steps until the tree is complete:

Selection step: based on the distances between taxa, choose two `active' taxa $i$ and $j$ to agglomerate. That is, connect them to a new taxon $(ij)$ representing their direct common ancestor. (Initially all taxa are `active'.)

Reduction step: remove i and j from the list of active taxa, and define new distances between the new active taxon (ij) and all remaining active taxa.

The two steps above are common to all agglomerative algorithms. For NJ, there are a number of equivalent (Gascuel 1994) ways to specify the two steps above (Saitou & Nei 1987, Studier & Keppler 1988). Here we describe the most efficient computationally (Studier & Keppler 1988). In the selection step, NJ agglomerates the taxa i and j that minimize

$$q_{ij} = (r-2)\delta_{ij} - \sum_{k=1}^{r} \delta_{ik} - \sum_{k=1}^{r} \delta_{jk}, \tag{4}$$

where the sums run on the set of remaining active taxa, and r is their number. As for the reduction step, the new distance between (ij) and any other taxon k is defined by

$$\delta_{(ij)k} = \tfrac{1}{2}\big(\delta_{ik} + \delta_{jk} - \delta_{ij}\big).$$

A third step, defining branch lengths for the reconstructed tree, is also usually described for NJ, but here we omit it for simplicity.

Different choices for the two steps above lead to other well-known agglomerative algorithms: single-linkage clustering (Sneath 1957), average-linkage clustering, also known as UPGMA (Sokal & Michener 1958), WPGMA (Sokal & Sneath 1963), ADDTREE (Sattath & Tversky 1977), UNJ (Gascuel 1997a), and BIONJ (Gascuel 1997b). The last method is a special case of the MVR approach (Gascuel 2000a), which adapts the reduction step above so as to account for the variances and covariances of the distance estimates δ_{ij}. It is similar in spirit to another

agglomerative algorithm, Weighbor (Bruno et al. 2000), which also modifies the selection criterion, and uses a different formula (Bulmer 1991) to evaluate the variances of the distances.

Of the algorithms above, the fastest are single-linkage clustering, UPGMA and WPGMA, as they manage to construct a tree in $O(n^2)$ time (Sibson 1973, Murtagh 1984, Gronau & Moran 2007). However they are only accurate when the distances are approximately ultrametric (see Glossary), and thus they are little used for phylogenetic inference, where the molecular clock is the exception rather than the rule. Apart from ADDTREE and MVR, which have a time complexity of $O(n^4)$, all other agglomerative algorithms mentioned above reconstruct a tree in $O(n^3)$ time. This includes NJ, where each of the $n-3$ selection steps is carried out in $O(r^2) = O(n^2)$ time via precalculation of the sums $\sum_{k=1}^{r} \delta_{ik}$ (Studier & Keppler 1988). One of the reasons for the continued success of NJ is the fact that it has long been considered to achieve a very good tradeoff between reconstruction accuracy and running times.

Recently, a lot of work has gone into crafting computationally efficient implementations of NJ, both in terms of running time and memory usage. These include QuickTree (Howe et al. 2002), QuickJoin (Mailund & Pedersen 2004), the bucket-based method of (Zaslavsky & Tatusova 2008), NINJA (Wheeler 2009), RapidNJ and ERapidNJ (Simonsen et al. 2011). A general idea is to speed up the selection of the pair of taxa that minimizes q_{ij} in equation (4), by limiting the search to a subset of pairs guaranteed to contain the best pair. Moreover, external (disk) memory is used to cope with large data sets (standard implementations can only deal with a few thousand taxa because of internal memory limitations). Although the worst-case time complexity remains $O(n^3)$, these approaches permit a dramatic improvement in efficiency, allowing the reconstruction of trees with more than 50,000 taxa in a few hours on a normal PC (Wheeler 2009, Simonsen et al. 2011).

Even faster NJ-like algorithms can be obtained by employing heuristics – instead of exact algorithms – to select the pair of taxa to join on the basis of q_{ij}. These include FastNJ (Elias & Lagergren 2009) and RelaxedNJ (Evans et al. 2006), which is implemented in Clearcut (Sheneman et al. 2006). In the case of FastNJ, the taxa to join are selected among a list of $O(n)$ pairs, implying a running time of $O(n^2)$. For both methods, some loss of accuracy is to be expected, as in general these approaches do not reconstruct the same tree as NJ. Despite this, like NJ, these methods are statistically consistent, as they are guaranteed to reconstruct a tree T when the input distances are additive with respect to T, or nearly additive (Atteson 1999, Elias & Lagergren 2009).

A question that puzzled phylogeneticists for some time is whether NJ is related to any of the optimization criteria we discussed above. It was often suggested that "NJ has some relation to ordinary least squares and some to minimum evolution, without being definable as an approximate algorithm for either" (Felsenstein 2004). These connections come from the fact that the original selection criterion q_{ij} can be obtained as a sum of OLS estimates for a certain subset of branch lengths (Saitou & Nei 1987, Gascuel 1994). However, if the good performance of NJ were due to its relation to OLS+ME, then we would expect that better (shorter) trees with respect to this criterion would also be phylogenetically more accurate than NJ trees, something that is actually contradicted by experience (Saitou and Imanishi 1989, Kumar 1996, Gascuel 2000b, Desper & Gascuel 2002).

More recently, it was shown that the real optimization criterion behind NJ is BME (Desper & Gascuel 2005, Gascuel & Steel 2006). To briefly illustrate this, we note that the formula $L(T)$ in equation (3) (which only applies to bifurcating trees) can be generalized to unresolved trees by replacing $2^{1-|\tau_{ij}|}$ by a factor $p(i \to j)$ expressing the probability of ending up in j when following a suitably defined random walk starting at i. The resulting generalized BME formula can then be seen as the guiding principle behind the selection step in NJ: at each of these steps, the agglomeration performed by NJ is the one that results in the tree with the smallest BME length.

Motivated by the observation that NJ can be seen as a greedy algorithm for BME, other methods guided by BME have been proposed (Desper & Gascuel 2002, Catanzaro et al. 2012). These methods differ from NJ in their choice for component C3, that is, the algorithm to seek the BME-optimal tree. FastME (Desper & Gascuel 2002, Lefort et al. 2015) implements heuristics including stepwise addition to construct an initial tree, and common tree topology rearrangements (NNI, SPR) to perform a local search in tree space. Although the running time of FastME is comparable to that of NJ, its reconstruction accuracy is superior (Desper & Gascuel 2002, 2004, Vinh & von Haeseler 2005), thus confirming the suitability of BME as an optimization principle in distance-based phylogenetics.

4.4 Beyond distances

An important computational bottleneck of all the methods we presented so far is that they require the initial estimation and storage of distances for all pairs of taxa, which takes $O(\ell n^2)$ time – assuming that distances are estimated from sequences of length ℓ – and $O(n^2)$ memory. It is intuitive, however, that not all distances are necessary to reconstruct a phylogeny, meaning that these bounds can be improved. Some distances – those with large variances – may even be misleading for tree inference. Moreover, for large datasets with hundreds of thousands taxa, reducing memory usage may be a necessity: just storing the entire distance matrix for 100K taxa typically requires 20 GB of memory, which may be problematic for many users.

In the last few years a number of approaches have been proposed to bypass the bottleneck above. The most widely used is FastTree (Price et al. 2009), whose strategy to construct an initial tree is inspired by NJ, but with two fundamental differences – one aimed at improving reconstruction accuracy, and the other for computational efficiency. We now briefly describe the main distinctive points of FastTree, as they underlie its good performance and thus its popularity.

The first difference with NJ and other classical distance-based methods is that FastTree stores *sequence profiles* for the active taxa (see Glossary) – instead of a distance matrix – and only computes the distance between two profiles if the corresponding pair is a candidate for joining. After each agglomeration, the sequence profile for the new node joining *i* and *j* is computed as the arithmetic average of the profiles for *i* and *j*. The pair of taxa to join is selected on the basis of a distance-based criterion formally similar to (4), but where distances are uncorrected and defined on the basis of the profiles stored for the active taxa. The second difference consists of maintaining a list of $O(\sqrt{n})$ `top-hits' for each taxon (i.e. putative closest neighbours), which is combined to the strategies of FastNJ (Elias & Lagergren 2009) and RelaxedNJ (Evans et al. 2006) to reduce the pairs of taxa to consider for agglomeration: at the end of its execution FastTree will only have considered $O(n^{1.5} \log n)$ pairs – as opposed to $O(n^3)$ for NJ, and $O(n^2)$ for FastNJ. The NJ-like reconstruction of an initial tree – with a claimed complexity of $O(\ell n^{1.5} \log n)$ time and $O(\ell n + n^{1.5})$ memory – is then followed by a local search based on common tree topology rearrangements (NNI, SPR), using techniques similar to those implemented in FastME (Desper & Gascuel 2002).

FastTree is faster and more memory-efficient than the distance-based methods discussed so far, and it can easily cope with data sets with hundreds of thousands of sequences. Given the central role of sequence profiles, it can be argued that FastTree is not a distance-based method. It shares ideas with with character-based methods such as parsimony, benefitting in particular from information about ancestral sequences – something that is not available to purely distance-based methodology.

Other methods that arguably lie beyond the frontier of distance methods – but which share with them several ideas and the same emphasis on computational efficiency – have recently been proposed by Brown, Truszkowski and collaborators (Truszkowski et al. 2012, Brown and Truszkowski 2012). Particularly promising as a basis for future methodology is LSHTree

(Brown and Truszkowski 2012), which like NJ proceeds in a bottom-up fashion by joining subtrees – although not necessarily at their roots. To do so, it uses `locality-sensitive hashing' exploiting ancestral sequence reconstructions, to rapidly find candidate pairs of close sequences to join. Just as FastTree, the running time of LSHTree is sub-quadratic.

5. Conclusion

Despite the simplicity of their approach, and the loss of information that is necessarily entailed by summarizing sequence data into a numerical matrix, distance-based methods are not just computationally efficient, but also remarkably accurate. The reconstruction accuracy of distance-based methods has been the subject of many simulation studies in the past (e.g., Saitou & Imanishi 1989, Kuhner & Felsenstein 1994, Kumar 1996, Gascuel 2000b, Nakhleh et al. 2002, Desper & Gascuel 2004). The general idea of these studies is to generate sequences using standard substitution models and known model trees, and then compare the trees reconstructed by a number of competing methods to the model trees employed to generate the sequences. From these works it transpires that although not comparable to that of likelihood-based methods, the reconstruction accuracy of distance-based methods is competitive with that of maximum parsimony (MP), with MP superior for trees with short branches, and inferior when the effect of multiple substitutions at the same site becomes important. The reason for this is largely intuitive: while distance-based methods naturally account (or `correct') for multiple substitutions in the way they estimate distances, MP does not even model branch lengths, leading to serious problems such as statistical inconsistency in extreme cases (Felsenstein 1978). Another important conclusion that empirical studies have helped to reach is the importance of improving the tree reconstructed initially, via topological rearrangements such as NNI and SPR (e.g., Vinh & von Haeseler 2005). This is why most modern tree reconstruction methods include this important step, as we have seen for FastME (Desper & Gascuel 2002) and FastTree (Price et al. 2009) in a distance-based context.

Another important advantage of distance-based methods is their versatility: they can be employed not just with sequence data, but in every context where pairwise comparisons are possible. For example, they have been used to infer phylogenies from morphological characters (Sokal & Michener 1958), immunological data (Sarich & Wilson 1967), gene frequencies (Cavalli-Sforza & Edwards 1967), DNA-DNA hybridization data (Sibley & Ahlquist 1984), and more recently from gene content (Snel et al. 1999) or gene order (Wang et al. 2006) within genomes. Another area where distance methods may prove useful is phylogenomics, where large collections of genomic alignments may be summarized into multiple distance matrices. Their combined analysis may provide an efficient alternative to traditional supertree and supermatrix approaches (Lapointe & Cucumel 1997, Criscuolo et al. 2006, Binet et al. 2015).

See also: Parsimony (205), Maximum likelihood (207), Models and Model Selection (209).

References

Atteson, K. (1999). The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, *25*(2-3), 251-278.

Binet, M., Gascuel, O., Scornavacca, C., Douzery, M., & Pardi, F. (2015). Fast and accurate branch length estimation for phylogenomic trees. *Submitted*.

Brown, D. G., & Truszkowski, J. (2012). Fast phylogenetic tree reconstruction using locality-sensitive hashing. In *Algorithms in bioinformatics* (pp. 14-29). Springer Berlin Heidelberg.

Bruno, W. J., Socci, N. D., & Halpern, A. L. (2000). Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular biology and evolution*, 17(1), 189-197.

Bryant, D., & Waddell, P. (1998). Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Molecular biology and evolution*, 15(10), 1346-1359.

Bulmer, M. (1991). Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular biology and evolution*, 8(6), 868-883.

Buneman, O. P. (1971). The recovery of trees from measures of dissimilarity. In: Hodson, F. R., Kendall D. G., Tautu, P. (eds.), *Mathematics in the archaeological and historical sciences*. Edinburgh University Press.

Catanzaro, D., Labbé, M., Pesenti, R., & Salazar-González, J. J. (2012). The balanced minimum evolution problem. *INFORMS journal on computing*, *24*(2), 276-294.

Cavalli-Sforza, L. L., & Edwards, A. W. (1967). Phylogenetic analysis. Models and estimation procedures. *American journal of human genetics*, 19(3 Pt 1), 233.

Chakraborty, R. (1977). Estimation of time of divergence from phylogenetic studies. *Canadian journal of genetics and cytology*, 19(2), 217-223.

Criscuolo, A., Berry, V., Douzery, E. J., & Gascuel, O. (2006). SDM: a fast distance-based approach for (super) tree building in phylogenomics. *Systematic biology*, *55*(5), 740-755.

Day, W. H. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of mathematical biology*, 49(4), 461-467.

Desper, R., & Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of computational biology*, 9(5), 687-705.

Desper, R., & Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular biology and evolution*, 21(3), 587-598.

Desper, R., & Gascuel, O. (2005). The minimum evolution distance-based approach to phylogenetic inference. In: Gascuel, O. (ed.) *Mathematics of evolution and phylogeny*, 1-32.

Elias, I., & Lagergren, J. (2009). Fast neighbor joining. *Theoretical computer science*, *410*(21), 1993-2000.

Evans, J., Sheneman, L., & Foster, J. (2006). Relaxed neighbor joining: a fast distance-based phylogenetic tree construction method. *Journal of molecular evolution*, *62*(6), 785-792.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic biology*, *27*(4), 401-410.

Felsenstein, J. (1997). An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic biology*, 46(1), 101-111.

Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer.

Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(3760), 279-284.

Gascuel, O. (1994). A note on Sattath and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Molecular biology and evolution*, 11(6), 961-963.

Gascuel, O. (1997a). Concerning the NJ algorithm and its unweighted version, UNJ. *Mathematical hierarchies and biology*, 37, 149-171.

Gascuel, O. (1997b). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and evolution*, *14*(7), 685-695.

Gascuel, O. (2000a). Data model and classification by trees: the minimum variance reduction (MVR) method. *Journal of classification*, *17*(1), 67-99.

Gascuel, O. (2000b). On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. *Molecular biology and evolution*, *17*(3), 401-405.

Gascuel, O., Bryant, D., & Denis, F. (2001). Strengths and limitations of the minimum evolution principle. *Systematic biology*, 621-627.

Gascuel, O., & Steel, M. (2006). Neighbor-joining revealed. *Molecular biology and evolution*, 23(11), 1997-2000.

Gronau, I., & Moran, S. (2007). Optimal implementations of UPGMA and other common clustering algorithms. *Information processing letters*, *104*(6), 205-210.

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5), 696-704.

Howe, K., Bateman, A., & Durbin, R. (2002). QuickTree: building huge Neighbour-Joining trees of protein sequences. *Bioinformatics*, *18*(11), 1546-1547.

Jukes, T. H., & Cantor, C. R. (1969). Evolution of protein molecules. In: Munro, H. N. (ed.), *Mammalian protein metabolism*, San Diego, CA: Academic Press, pp. 21–132.

Kidd, K. K., & Sgaramella-Zonta, L. A. (1971). Phylogenetic analysis: concepts and methods. *American journal of human genetics*, 23(3), 235.

Kuhner, M. K., & Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular biology and evolution*, 11(3), 459-468.

Kumar, S. (1996). A stepwise algorithm for finding minimum evolution trees. *Molecular biology and evolution*, *13*(4), 584-593.

Lapointe, F. J., & Cucumel, G. (1997). The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic biology*, *46*(2), 306-312.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., ... & Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21), 2947-2948.

Lawson, C. L., & Hanson, R. J. (1974). *Solving least squares problems*. Prentice-hall.

Lefort, V., Desper, R., & Gascuel, O. (2015). FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular biology and evolution*, in press.

Mailund, T., & Pedersen, C. N. (2004). QuickJoin — fast neighbour joining tree reconstruction. *Bioinformatics*, *20*(17), 3261-3262.

Mihaescu, R., & Pachter, L. (2008). Combinatorics of least-squares trees. *Proceedings of the national academy of sciences USA*, 105(36), 13206-13211.

Murtagh, F. (1984). Complexities of hierarchic clustering algorithms: State of the art. *Computational statistics quarterly*, *1*(2), 101-113.

Nakhleh, L., Moret, B. M., Roshan, U., John, K. S., & Warnow, T. (2002). The accuracy of fast phylogenetic methods for large datasets. In *Proc. 7th Pacific Symp. Biocomputing PSB 2002* (pp. 211-222).

Pardi, F., & Gascuel, O. (2012). Combinatorics of distance-based tree inference. *Proceedings of the national academy of sciences USA*, *109*(41), 16443-16448.

Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *Journal of molecular evolution*, *51*(1), 41-47.

Price, M. N., Dehal, P. S., and A. P. Arkin (2009) FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution,* 26: 1641-1650.

Rzhetsky, A., & Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular biology and evolution*, 10(5), 1073-1095.

Saitou, N., & Imanishi, T. (1989). Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Molecular biology and evolution*, 6(5), 514-525.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4), 406-425.

Sarich, V. M., & Wilson, A. C. (1967). Immunological Time Scale for Hominid Evolution. *Science*, 158(3805), 1200-1203.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*(3), 319-345.

Semple, C., & Steel, M. (2004). Cyclic permutations and evolutionary trees. *Advances in applied mathematics*, 32:669–80.

Sheneman, L., Evans, J., & Foster, J. A. (2006). Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics*, *22*(22), 2823-2824.

Sibley, C. G., & Ahlquist, J. E. (1984). The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *Journal of molecular evolution*, *20*(1), 2-15.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *The computer journal*, *16*(1), 30-34.

Simões Pereira, J.M.S. (1969). A note on tree realizability of a distance matrix. *Journal of combinatorial theory*, 6(B), 303-310.

Simonsen, M., Mailund, T., & Pedersen, C. N. (2011). Inference of large phylogenies using neighbour-joining. In *biomedical engineering systems and technologies* (pp. 334-344). Springer Berlin Heidelberg.

Sokal R., & Michener C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas science bulletin,* 38: 1409–1438.

Sokal, R. R., & Sneath, P. H. (1963). *Principles of numerical taxonomy*. W.H. Freeman.

Sneath, P. H. (1957). The application of computers to taxonomy. *Journal of general microbiology*, *17*(1), 201-226.

Snel, B., Bork, P., & Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nature genetics*, *21*(1), 108-110.

Studier, J. A., & Keppler, K. J. (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular biology and evolution*, *5*(6), 729-731.

Swofford, D. L., Olsen, G. J., Waddell, P. J., & Hillis, D. M. (1990). Phylogeny reconstruction. In: Hiillis, D. M., Moritz, C., & Mable, B.K. (eds.), *Molecular systematics*, 3, 407-514.

Swofford, D. L. (2003). PAUP*. *Phylogenetic Analysis Using Parsimony* (* *and Other Methods*). Version 4, Sunderland, MA: Sinauer Associates.

Truszkowski, J., Hao, Y., & Brown, D. G. (2012). Towards a practical O (n log n) phylogeny algorithm. *Algorithms for molecular biology*, *7*(1), 32.

Vach, W. (1989). Least squares approximation of additive trees. In Opitz, O. (Ed.), *Conceptual and Numerical Analysis of Data*, Berlin: Springer-Verlag, pp. 230–238.

Vinh, L. S., & von Haeseler, A. (2005). Shortest triplet clustering: reconstructing large phylogenies using representative sets. *BMC bioinformatics*, *6*(1), 92.

Wang, L. S., Warnow, T., Moret, B. M., Jansen, R. K., & Raubeson, L. A. (2006). Distance-based genome rearrangement phylogeny. *Journal of molecular evolution*, *63*(4), 473-483.

Wheeler, T. J. (2009). Large-scale neighbor-joining with NINJA. In S.L. Salzberg and T. Warnow (Eds.), *Proceedings of the 9th workshop on algorithms in bioinformatics. WABI 2009*, pp. 375-389. Springer, Berlin.

Yang, Z. (2006). *Computational molecular evolution*. Oxford University Press.

Zaslavsky, L., & Tatusova, T. A. (2008). Accelerating the neighbor-joining algorithm using the adaptive bucket data structure. In *Bioinformatics research and applications*, Berlin Heidelberg: Springer, pp. 122-133.

Zaretskii, K. A. (1965). Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspekhi Matematicheskikh Nauk*, 20(6), 90-92 (in Russian).