Distance-Based Phylogeny Reconstruction: Safety and Edge Radius

Olivier Gascuel, Fabio Pardi, Jakub Truszkowski

Distance-based Phylogeny Reconstruction: Safety and Edge Radius

O. Gascuel[1], F. Pardi[1], J. Truszkowski[2,3]

[1] Institut de Biologie Computationnelle, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), UMR 5506, CNRS & Université de Montpellier, Bâtiment 5, 860 rue de Saint Priest, 34095 Montpellier cedex 5, France

[2] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom

[3] Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge CB2 0RE, United Kingdom

Years and Authors of Summarized Original Work

1999; Atteson

2005; Elias, Lagergren

2006; Dai, Xu, Zhu

2010; Pardi, Guillemot, Gascuel

2013; Bordewich, Mihaescu

Keywords

phylogeny reconstruction, distance methods, performance analysis, robustness, safety radius approach, optimal radius

Problem Definition

A phylogeny is an evolutionary tree tracing the shared history, including common ancestors, of a set of extant species or "taxa". Phylogenies are increasingly reconstructed on the basis of molecular data (DNA and protein sequences) using statistical techniques such as likelihood and Bayesian methods. Algorithmically, these techniques suffer from the discrete nature of tree topology space. Since the number of tree topologies increases exponentially as a function of the number of taxa, and each topology requires a separate likelihood calculation, it is important to restrict the search space and to design efficient heuristics. Distance methods for phylogeny reconstruction serve this purpose by inferring trees in a fraction of the time required for the more statistically rigorous methods. Distance methods also provide fairly accurate starting trees to be further refined by more sophisticated methods. Moreover, the input to a distance method is the matrix of pairwise evolutionary distances among taxa, which are estimated by maximum likelihood, so that distance methods also have sound statistical justifications.

Mathematically, a phylogenetic tree is a triple $T = (V, E, l)$ where V is the set of nodes (extant taxa correspond to leaves, ancestors to internal nodes), E is the set of edges (branches) representing relations of descent, and l is a function that assigns positive lengths to each edge in E, representing a measure of evolutionary divergence, for example in terms of time, or amount of change between DNA or protein sequences. Any phylogenetic tree T defines a metric D_T on its leaf set L: let $P_T(u,v)$ define the unique path through T from u to v, then the distance from u to v is set to $D_T(u, v) = \sum_{e \in P_T(u,v)} l(e)$.

Distance methods for phylogeny reconstruction rely on the fundamental result [21] that the map $T \rightarrow D_T$ is reversible; i.e., a tree T can be reconstructed from its tree metric, a problem that can be solved in $O(n \log n)$ time [13]. However, in practice D_T is not known, and one must use molecular sequence data to estimate a distance matrix D that approximates D_T [9]. As the amount of sequence data increases, D can be assumed to converge to D_T. A minimal requirement for any distance method is *consistency*: for any tree T, and for distance matrices D "close enough" to D_T, the algorithm should output a tree with the same topology as T (that is, with the same underlying graph (V,E)). The present chapter deals with the question of when any distance algorithm for phylogeny reconstruction can be guaranteed to output the correct phylogeny as a function of the divergence between D and D_T. Atteson [1] demonstrated that this question can be precisely answered for Neighbor Joining (NJ) [17], one of the most cited algorithms in computational biology (with more than 35,000 citations up to 2014), and a number of NJ's variants.

The Neighbor Joining (NJ) algorithm of Saitou and Nei (1987)

NJ is *agglomerative*: it works by using the input matrix D to identify a pair of taxa $x, y \in L$ that are neighbors in T, i.e. there exists a node $u \in V$ such that $\{(u,x), (u,y)\} \subset E$. Then, the algorithm creates a node c that is connected to x and y, extends the distance matrix to c, and solves the reduced problem on $L \cup \{c\} \setminus \{x,y\}$. The pair (x,y) is chosen to minimize the following sum:

$$S_D(x,y) = (|L| - 2) \cdot D(x,y) - \sum_{z \in L} (D(z,x) + D(z,y)).$$

The soundness of NJ is based on the observation that, if $D = D_T$ for a tree T, the value $S_D(x,y)$ will be minimized for a pair (x,y) that are neighbors in T.

Balanced Minimum Evolution and algorithms inspired by it

A number of papers (reviewed in [11]) have been dedicated to the various interpretations and properties of the S_D criterion. One of these interpretations consists in observing that agglomerating the pair of nodes that minimizes S_D is equivalent to choosing, among all the trees that can be obtained in this way, the one that minimizes a simple linear formula [15] to calculate the length of a tree from the distances between its leaves [11], thus connecting distance and parsimony methods [9]. As the optimization principle seeking the tree that minimizes this formula has been named Balanced Minimum Evolution (BME) [6], NJ can then be seen as a greedy algorithm for BME.

This remarkable connection between NJ and BME naturally spurred the proposal of alternative algorithms for BME. One of these, GreedyBME, consists of iteratively adding taxa to a tree so that, at each step, the resulting tree is the one that minimizes BME among all the binary trees that can be obtained in this way [6]. More involved algorithms can be obtained by combining a simple tree construction algorithm such as NJ or GreedyBME, with a local search

based on the traditional tree rearrangements used in phylogenetics [9], such as *nearest-neighbor interchange* (NNI) or *subtree pruning and regrafting* (SPR).

The Fast Neighbor Joining (FNJ) Algorithm of Elias and Lagergren (2005)

Standard implementations of NJ require $O(n^3)$ computations, where n is the number of taxa in the data set. Since a distance matrix only has n^2 entries, many attempts have been made to construct a distance algorithm that would only require $O(n^2)$ computations while retaining the accuracy of NJ. To this end, one of the most interesting results is the Fast Neighbor Joining (FNJ) algorithm of Elias and Lagergren [7].

Most of the computation of NJ is used in the recalculations of the sums $S_D(x,y)$ after each agglomeration step. Although each recalculation can be performed in constant time, and although it is not necessary to consider all pairs of taxa (x,y) in order to find the one that minimizes this sum [19], the number of pairs to consider remains, in the worst case, $O(k^2)$ when k nodes are left to agglomerate. Thus, summing over k, $O(n^3)$ computations are required in all.

Elias and Lagergren take a related approach to agglomeration, which does not exhaustively seek the minimum value of $S_D(x,y)$ at each step, but instead uses a heuristic to maintain a list of candidates of "visible pairs" (x,y) for agglomeration. At the $(n - k)^{\text{th}}$ step, when two neighbors are agglomerated from a k-taxa tree to form a $(k - 1)$-taxa tree, FNJ has a list of $O(k)$ visible pairs for which $S_D(x,y)$ is calculated. The pair joined is selected from this list. By trimming the number of pairs considered, Elias and Lagergren achieved an algorithm which requires only $O(n^2)$ computations. Other similar improvements to Neighbor Joining have also been proposed in recent years [8,12,19].

Safety radius performance analysis (Atteson 1999)

In order to provide accuracy guarantees for distance-based algorithms, Atteson [1] tackled the following question: if D is a distance matrix that approximates a tree metric D_T, can one have some confidence in algorithm's ability to reconstruct T, or parts of it, given D, based on some measure of the distance between D and D_T? For two matrices, D_1 and D_2, the L_∞ distance between them is defined by $||D_1 - D_2||_\infty = \max_{i,j} |D_1(i, j) - D_2(i, j)|$. Moreover, let $\mu(T)$ denote the length of the shortest internal edge of a tree T. This is an important quantity, as short branches in a phylogeny are difficult to resolve, because of the relatively few (if any) molecular changes occurring on a short branch.

The *safety radius* of an algorithm A is then the greatest value of r with the property that: given any phylogeny T, and any distance matrix D satisfying $||D - D_T||_\infty < r \cdot \mu(T)$, A will return a tree $\hat{T}$ with the same topology as T. Similarly, the *edge radius* of A is the greatest value of r, for which the presence in $\hat{T}$ of an edge $e \in E$ is guaranteed whenever $||D - D_T||_\infty < r \cdot l(e)$. As an easy consequence of these definitions, the safety radius is always at least as large as the edge radius. Moreover, both the safety radius and the edge radius can also be attributed to an optimization principle, assuming an exact optimization algorithm.

Key Results

Atteson [1] proved the following theorems.

Theorem 1: The safety radius of NJ is ½.

Theorem 2: The largest possible safety radius for any algorithm is ½.

Indeed, given any μ, one can find two different trees T_1, T_2 and a distance matrix D such that $\mu = \mu(T_1) = \mu(T_2)$, and $\|D - D_{T_1}\|_\infty = \mu/2 = \|D - D_{T_2}\|_\infty$. Since D is equidistant from two distinct tree metrics, no algorithm could assign it to the "closest" tree.

In their presentation of FNJ, Elias and Lagergren updated Atteson's results for their algorithm. They showed

Theorem 3: The safety radius of FNJ is ½.

An insight on the above results on neighbor-joining-type algorithms is provided by the fact that the optimization principle they are linked to, BME, has itself safety radius ½ [14]. A simple consequence of this [14] is the fact that also GreedyBME has safety radius ½, a result first proven by Shigezumi [18]. Finally, performing a local search guided by BME and based on SPR leads to an algorithm with safety radius greater or equal to ⅓, regardless of the method used to construct the initial tree [2].

The edge radius of a number of algorithms has also been studied. As conjectured by Atteson [1] and formally proven by Dai et al. [5], the edge radius of NJ is ¼. Interestingly, other heuristics, related to NJ via the principle they seek to optimize (BME), perform better than NJ in terms of edge radius: GreedyBME has edge radius ⅓ [3]; moreover, building an initial tree with GreedyBME and then performing a local search guided by BME and based on NNI or SPR operations, constitutes an algorithm with edge radius ⅓ [3].

Finally we note that the safety radius framework has also been applied to the ultrametric setting where the correct tree T is rooted and all tree leaves are at the same distance from the root [10]. These trees are called "molecular clock" trees in phylogenetics and "indexed hierarchies" in data analysis. In this setting, the optimal safety radius is equal to 1 (instead of ½) and a number of standard algorithms (e.g. UPGMA, with time complexity in $O(n^2)$) have a safety radius of 1.

Open Problems

With increasing amounts of sequence data becoming available for an increasing number of species, distance algorithms such as NJ should be useful for quite some time. Currently, the bottleneck in the process of building phylogenies is estimating distances, rather than exploring tree topologies. Two algorithms were recently developed to reconstruct trees from incomplete

distance matrices. These algorithms use character information as well as distances, and hence cannot be categorized as pure distance methods.

FastTree [16] is an NJ-like heuristic that avoids computing the full distance matrix. For each taxon, FastTree computes the distances to $O(\sqrt{n})$ close neighbours. FastTree also uses sequence profiles to approximate $S_D(x,y)$ values in constant time. The overall algorithm takes $O(san\sqrt{n}\log n)$ time and $O(san + n\sqrt{n})$ memory, where s is the length of the input sequences and a is their alphabet size. FastTree has been shown to be highly accurate with simulated data [16], but no formal guarantee has yet been shown for this algorithm.

The only known $o(n^2)$ algorithm with theoretical guarantees is LSHTree [4]. It uses locality-sensitive hashing to rapidly find candidate pairs of close sequences for merging. After each merge, LSHTree reconstructs ancestral sequences at new internal nodes to ensure that a close pair of sequences can be found at each iteration. LSHTree is guaranteed to reconstruct the correct tree from sequences of logarithmic length under a Markov model of sequence evolution. The exact running time of LSHTree depends on the branch lengths.

As we have shown, a number of distance-based tree building algorithms have been analyzed in the safety radius framework. However, computer simulations (e.g. [6,7]) have shown that not all algorithms with optimal safety radius achieve the same accuracy: for example, NJ is slightly more accurate than FNJ (both having safety radius = ½), but is beaten by heuristics based on NNI or SPR moves (with demonstrated safety radius >= ⅓, but possibly = ½). Moreover, some well-established methods (e.g. based on least-squares [10,20]) have safety radius converging to 0 when the number of taxa increases, which contradicts the common practice. These experimental observations indicate that the safety radius approach should be sharpened to provide better theoretical analysis of method performance [22]. In particular, the choice of the L_∞ norm to measure the error in a distance matrix seems to have little statistical or biological justification.

An alternative analysis framework, strictly linked to the one presented here, is the one seeking to estimate the minimum sequence length required for accurate reconstruction of the correct tree. It is discussed in a separate entry of this Encyclopaedia [A].

Cross References

[A] Distance-Based Phylogeny Reconstruction (Fast-Convergence)

Recommended Reading

[1] K. ATTESON, *The performance of neighbor-joining methods of phylogenetic reconstruction*, Algorithmica, 25, 251-278 (1999)

[2] M. BORDEWICH, O. GASCUEL, K.T. HUBER, AND V. MOULTON. *Consistency of topological moves based on the balanced minimum evolution principle of phylogenetic inference.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 6, 110-117 (2009)

[3] M. BORDEWICH AND R. MIHAESCU. *Accuracy guarantees for phylogeny reconstruction algorithms based on balanced minimum evolution*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10, 576-583 (2013)

[4] D.G. BROWN AND J. TRUSZKOWSKI. *Fast phylogenetic tree reconstruction using locality-sensitive hashing.* Algorithms in Bioinformatics. Springer Berlin Heidelberg, pp. 14-29 (2012)

[5] W. DAI, Y. XU, AND B. ZHU. *On the edge l∞ radius of Saitou and Nei's method for phylogenetic reconstruction*. Theoretical Computer Science, 369, 448-455 (2006).

[6] R. DESPER AND O. GASCUEL, *Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum- Evolution Principle*, Journal of Computational Biology, 9, 687-706 (2002)

[7] I. ELIAS AND J. LAGERGREN, *Fast Neighbor Joining.,* In: Proceedings of the 32nd International Colloquium on Automata, Languages, and Programming (ICALP), pp. 1263-1274 (2005)

[8] J. EVANS, L. SHENEMAN, AND J. FOSTER. *Relaxed neighbor joining: a fast distance-based phylogenetic tree construction method*. Journal of Molecular Evolution, 62, 785-792 (2006)

[9] J. FELSENSTEIN, *Inferring Phylogenies*, Sinauer Associates, Sunderland, Massachusetts (2004).

[10] O. GASCUEL AND A. MCKENZIE, *Performance Analysis of Hierarchical Clustering Algorithms*, Journal of Classification, 21, 3-18 (2004)

[11] O. GASCUEL AND M. STEEL, *Neighbor-Joining Revealed*. Molecular Biology and Evolution, 23, 1997-2000 (2006)

[12] I. GRONAU AND S. MORAN. *Neighbor joining algorithms for inferring phylogenies via lca distances*. Journal of Computational Biology, 14, 1-15 (2007)

[13] J. HEIN. *An optimal algorithm to reconstruct trees from additive distance data.* Bulletin of Mathematical Biology, 51, 597-603 (1989)

[14] F. PARDI, S. GUILLEMOT, AND O. GASCUEL. *Robustness of phylogenetic inference based on minimum evolution.* Bulletin of Mathematical Biology, 72, 1820-1839 (2010)

[15] Y. PAUPLIN, *Direct calculation of a tree length using a distance matrix*. Journal of Molecular Evolution, 51, 41-47 (2000)

[16] M.N. PRICE, P.S. DEHAL, AND A. P. ARKIN. *FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.* Molecular Biology and Evolution 26, 1641-1650 (2009)

[17] N. SAITOU AND M. NEI, *The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees,* Molecular Biology and Evolution, 4, 406-425 (1987)

[18] T. SHIGEZUMI. *Robustness of greedy type minimum evolution algorithms*. Computational Science–ICCS, 815-821 (2006)

[19] M. SIMONSEN, T. MAILUND AND C.N.S. PEDERSEN. *Inference of large phylogenies using neighbour-joining*. In: A Fred, J Felipe, H Gamboa (eds.) Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science Volume 127, pp. 334-344. Springer Berlin Heidelberg (2011)

[20] S. WILLSON, *Minimum evolution using ordinary least-squares is less robust than neighbor-joining.* Bulletin of Mathematical Biology 67, 261-279 (2005)

[21] K. ZARESTKII, *Reconstructing a tree from the distances between its leaves.* (In Russian) Uspehi Mathematicheskikh Nauk 20, 90-92 (1965)

[22] O. GASCUEL, M. STEEL. *A 'stochastic safety radius' for distance-based tree reconstruction*. Algorithmica, DOI 10.1007/s00453-015-0005-y (2015)