Distance-based Phylogeny Reconstruction (Optimal Radius), **1999, Atteson; 2005, Elias, Lagergren

Richard Desper, University College London,
http://abacus.gene.ucl.ac.uk/rick/rick.html

Olivier Gascuel, LIRMM, Centre Nationale de la Recherche Scientifique, France,
http://www.lirmm.fr/~gascuel

Synonyms and Index Terms: phylogeny reconstruction, distance methods, performance analysis, robustness, safety radius approach, optimal radius

1. Problem Definition

A phylogeny is an evolutionary tree tracing the shared history, including common ancestors, of a set of extant taxa. Phylogenies have been historically reconstructed using character-based (parsimony) methods, but in recent years the advent of DNA sequencing, along with the development of large databases of molecular data, has led to more involved methods. Sophisticated techniques such as likelihood and Bayesian methods are used to estimate phylogenies with sound statistical justifications. However, these statistical 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 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. They allow dealing with thousands of taxa, while the current implementations of statistical approaches are limited to a few hundreds, and distance methods also provide fairly accurate starting trees to be further refined by more sophisticated methods. Moreover, the input of distance methods 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 representing extant taxa and ancestral species, E is the set of edges (branches), and l is a function that assigns positive lengths to each edge in E. Evolution proceeds through the tree structure as a stochastic process with a finite state

space corresponding to the DNA bases or amino acids present in the DNA or protein sequences, respectively.

Any phylogenetic tree *T* defines a metric D_T on its leaf set *L*(*T*) : 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 observation [13] that the map $T \to D_T$ is reversible; i.e., a tree *T* can be reconstructed from its tree metric. While in practice D_T is not known, by using models of evolution (e.g. [10], reviewed in [5]) one can use molecular sequence data to estimate a distance matrix *D* that approximates D_T. As the amount of sequence data increases, the consistency of the various models of sequence evolution implies that *D* should converge to D_T. Thus for a distance method to be *consistent*, it is necessary that for any tree *T*, and for distance matrices *D* "close enough" to D_T, the algorithm will output *T*.

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 the metric underlying the true phylogeny and the metric estimated from the data. Atteson [1] demonstrated that this consistency can be shown for Neighbor Joining (NJ) [11], the most popular distance method, 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 $\{(u,x),(u,y)\} \subset E$. The algorithm creates a node *c* that is connected to *x* and *y*, extends the distance matrix to *c*, and then solves the reduced problem on $L \cup \{c\} \setminus \{x,y\}$. The pair (*x*,*y*) is chosen to minimize the following sum:

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

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*. A number of

papers (reviewed in [8]) have been dedicated to the various interpretations and properties of the S_D criterion.

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

NJ requires $\Omega(n^3)$ 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, the best result so far is the Fast Neighbor Joining (FNJ) algorithm of Elias and Lagergren [4].

Most of the computation of NJ is used in the re-calculations of the sums $S_D(x, y)$ after each agglomeration step. Although each recalculation can be performed in constant time, the number of such pairs is $\Omega(k^2)$ when *k* nodes are left to agglomerate, and thus, summing over *k*, $\Omega(n^3)$ 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 $(n\text{-}k)^{\text{th}}$ 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.

Safety radius-based performance analysis (Atteson 1999)

Short branches in a phylogeny are difficult to resolve, especially when they are nested deep within a tree, because relatively few mutations occurring on a short branch as opposed to on much longer pendant branches, which hides phylogenetic signal. One is faced with the choice between leaving certain evolutionary relationships unresolved (i.e., having an internal node with degree > 3), or examining when confidence can be had in the resolution of a short internal edge.

A natural formulation [9] of this question is: how long must be molecular sequences before one can have confidence in algorithm's ability to reconstruct *T* accurately? An alternative formulation [1] appropriate for distance methods: 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* 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*.

The latter formulation leads to a definition: The *safety radius* of an algorithm $\mathcal{A}$ is the greatest value of *r* with the property that: given any phylogeny *T*, and any distance matrix *D* satisfying $\|D - D_T\|_\infty < r \cdot \mu(T)$, $\mathcal{A}$ will return the tree *T*.

2. Key Results

Atteson [1] answered the second question affirmatively, with two theorems.

Theorem 1: The safety radius of NJ is $1/2$.

Theorem 2: For no distance algorithm $\mathcal{A}$ is the safety radius of $\mathcal{A}$ greater than $1/2$.

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 an optimally fast version of the NJ algorithm, Elias and Lagergren updated Atteson's results for the FNJ algorithm. They showed

Theorem 3: The safety radius of FNJ is $1/2$.

Elias and Lagergren showed that if *D* is a distance matrix and D_T is a tree metric with $\|D - D_T\|_\infty < \mu(T)/2$, then FNJ will output the same tree (*T*) as NJ.

3. Applications

Phylogeny is a quite active field within evolutionary biology and bioinformatics. As more proteins and DNA sequences become available, the need for fast and accurate phylogeny estimation algorithms is ever increasing as phylogeny not only serves to reconstruct species history but also to decipher genomes. To date, NJ remains one of

the most popular algorithms for phylogeny building, and is by far the most popular of the distance methods, with well over 1000 citations per year.

4. 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 not the problem of searching topology space, but rather the problem of building distance matrices. The brute force method to build a distance matrix on n taxa from sequences with l positions requires $\Omega(ln^2)$ computations, and typically $l >> n$. Elias and Lagergren proposed an $\Omega(ln^{1.376})$ algorithm based on Hamming distance and matrix calculations. However, this algorithm only applies to over-simple distance estimators [10]. Extending this result to more realistic models would be a great advance.

A number of distance-based tree building algorithms have been analyzed in the safety radius framework. Atteson [1] dealt with a large class of neighbor joining-like algorithms, and Gascuel and McKenzie [7] studied the ultrametric setting where the correct tree T is rooted and all tree leaves are at the same distance from the root. Such trees are very common; they 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 $1/2$) and a number of standard algorithms (e.g. UPGMA, with time complexity in $O(n^2)$) have a safety radius of 1. However, experimental studies (see below) showed that not all algorithms with optimal safety radius achieve the same accuracy, indicating that the safety radius approach should be sharpened to provide better theoretical analysis of method performance.

5. Experimental Results

Computer simulation is the most standard way to assess algorithm accuracy in phylogenetics. A tree is randomly generated as well as a sequence at tree root, whose evolution is simulated along the tree edges. A reconstruction algorithm is tested using the sequences observed at the tree leaves, thus mimicking the phylogenetic task. Various measures exist to compare the correct and the inferred trees, and algorithm performance is assessed as the average measure over repeated experiments. Elias and Lagergren [4] showed that FNJ (in $O(n^2)$) is just slightly outperformed by NJ (in

$O(n^3)$), while numerous simulations (e.g. [3, 12]) indicated that NJ is beaten by more recent algorithms (all in $O(n^3)$ or less), namely BioNJ [6], WEIGHBOR [2], FastME [3] and STC [12].

6. Data Sets

A large number of data sets is stored by the TreeBASE project, at http://www.treebase.org/.

7. URL to Code

For a list of leading phylogeny packages, see Joseph Felsenstein's website at http://evolution.genetics.washington.edu/phylip/software.html.

8. Cross References

173, 175, 180, 194, 200, 496, 532

9. Recommended Reading

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

[2] W.J. BRUNO, N.D. SOCCI, AND A. L. HALPERN, *Weighted Neighbor Joining: A Likelihood-Based Approach to Distance-Based Phylogeny Reconstruction*, Molecular Biology and Evolution, 17 (2000), pp. 189-197.

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

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

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

[6] O. GASCUEL, *BIONJ: an Improved Version of the NJ Algorithm Based on a Simple Model of Sequence Data*, Molecular Biology and Evolution, 14 (1997), pp. 685-695.

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

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

[9] D.H. HUSON, S. NETTLES, AND T. WARNOW, *Disk-covering, a fast-converging method for phylogenetic tree reconstruction*, Journal of Computational Biology, 6 (1999), pp. 369-386.

[10] T.H. JUKES AND C.R. CANTOR, *Evolution of Protein Molecules*, in Mammalian Protein Metabolism, H.N. Munro ed., Academic Press, New York, 1969, pp. 21-132.

[11] N. SAITOU AND M. NEI, *The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees*, Molecular Biology and Evolution, 4 (1987), pp. 406-425.
[12] L. S. VINH AND A. VON HAESELER, *Shortest triplet clustering: reconstructing large phylogenies using representative sets*, BMC Bioinformatics, 6 (2005), pp. 92.
[13] K. ZARESTKII, *Reconstructing a tree from the distances between its leaves*, Uspehi Mathematicheskikh Nauk, 20 (1965), pp. 90-92 (in Russian).