REVIEW

Neighbor-Joining Revealed

Olivier Gascuel* and Mike Steel†

*LIRMM, Montpellier, France and †Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

It is nearly 20 years since the landmark paper (Saitou and Nei 1987) in *Molecular Biology and Evolution* introducing Neighbor-Joining (NJ). The method has become the most widely used method for building phylogenetic trees from distances, and the original paper has been cited about 13,000 times (Science Citation Index). Yet the question "what does the NJ method seek to do?" has until recently proved somewhat elusive, leading to some imprecise claims and misunderstanding. However, a rigorous answer to this question has recently been provided by further mathematical investigation, and the purpose of this note is to highlight these results and their significance for interpreting NJ. The origins of this story lie in a paper by Pauplin (2000) though its continuation has unfolded in more mathematically inclined literature. Our aim here is to make these findings more widely accessible.

First we review briefly the Neighbor-Joining (NJ) method and outline what it does "not" do. NJ builds a tree from a matrix of pairwise evolutionary distances relating the set of taxa being studied. The distance between any taxon pair i and j is denoted as $d(i, j)$ and can be obtained from sequence data by a variety of approaches, for example, using Kimura's (1980) 2-parameter estimate. NJ iteratively selects a taxon pair, builds a new subtree, and agglomerates the pair of selected taxa to reduce the taxon set by one (fig. 1). Pair selection is based on choosing the pair i, j that minimizes the following Q criterion:

$$Q(i,j) = (r-2)d(i,j) - \sum_{k=1}^{r} d(i,k) - \sum_{k=1}^{r} d(j,k), \quad (1)$$

where r is the current number of taxa and the sums run on the taxon set. This formula is that of Studier and Keppler (1988), which we showed (Gascuel 1994) to be essentially equivalent to that of Saitou and Nei. Let f, g be the selected pair (fig. 1). NJ estimates the length of the branch (f, u) using

$$d(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(r-2)}\left[\sum_{k=1}^{r} d(f,k) - \sum_{k=1}^{r} d(g,k)\right], \quad (2)$$

and $d(g, u)$ is obtained by symmetry. Finally, NJ replaces f and g by u in the distance matrix, using the reduction formula:

$$d(u,k) = \frac{1}{2}[d(f,k) - d(f,u)] + \frac{1}{2}[d(g,k) - d(g,u)]. \quad (3)$$

Again, this formula is that of Studier and Keppler, not that of Saitou and Nei, but they are equivalent and the 2 NJ versions always reconstruct the same tree, both in terms of topology and branch lengths (Gascuel 1994).

Even though NJ behaved well, both with simulated and real data (Saitou and Nei 1987; Saitou and Imanishi

1989; Kuhner and Felsenstein 1994), mathematical foundations of these 3 formulas only became clear over several years, following mathematical investigations. The 2 main questions related to consistency (i.e., whether NJ correctly reconstructs a tree when the distances fit perfectly on that tree) and to the criterion that NJ optimizes. The consistency proofs of Saitou and Nei and Studier and Keppler were contested by Mirkin (1996) but fixed by Gascuel (1997) and Atteson (1999). The latter even proved a stronger result, showing that NJ still reconstructs the correct tree when the distance matrix is perturbed by small noise and that NJ is optimal regarding tolerable noise amplitude. Recently, Bryant (2005) provided a simple and elegant proof of NJ consistency, and so the first question has been well and truly laid to rest.

The second question was that of the criterion being optimized and of its biological relevance. Indeed, most phylogenetic methods are based on explicit criteria, for example, parsimony or maximum likelihood, and understanding properties of these criteria is a central issue, which is independent of (and should precede) the design of optimization algorithms. Saitou and Nei (see also Saitou 1996; Nei and Kumar 2000) showed that the selection criterion (1) is related to minimization of tree length, which is conceptually close to parsimony. Assuming that the current taxon set (a, b, c, d, e, f, and g in fig. 1a) only contains original taxa (i.e., not resulting from previous agglomerations, as is the case with a, f, and g), they proved that selection criterion (1) is equivalent to minimizing the ordinary (unweighted) least-squares (OLS) length estimate (Felsenstein 2004, p 148) of the agglomerated tree (as shown in fig. 1b, but removing dashed subtrees). Accordingly, they called this approach the minimum evolution (ME) principle, and the criterion being minimized by NJ was thought to be the OLS tree length estimate. However, this result was not fully convincing. First, the biological meaning and consistency of this ME + OLS principle was not established at the time of its publication. This was partly solved by Rzhetsky and Nei (1993) who demonstrated ME + OLS consistency; but the biological relevance of the OLS component was still questionable because it does not account for the high variance of the large evolutionary distances, as weighted least-squares (WLS) approaches do (Fitch and Margoliash 1967). Second, the result of Saitou and Nei applies to the first NJ step where we only have original taxa in the taxon set but
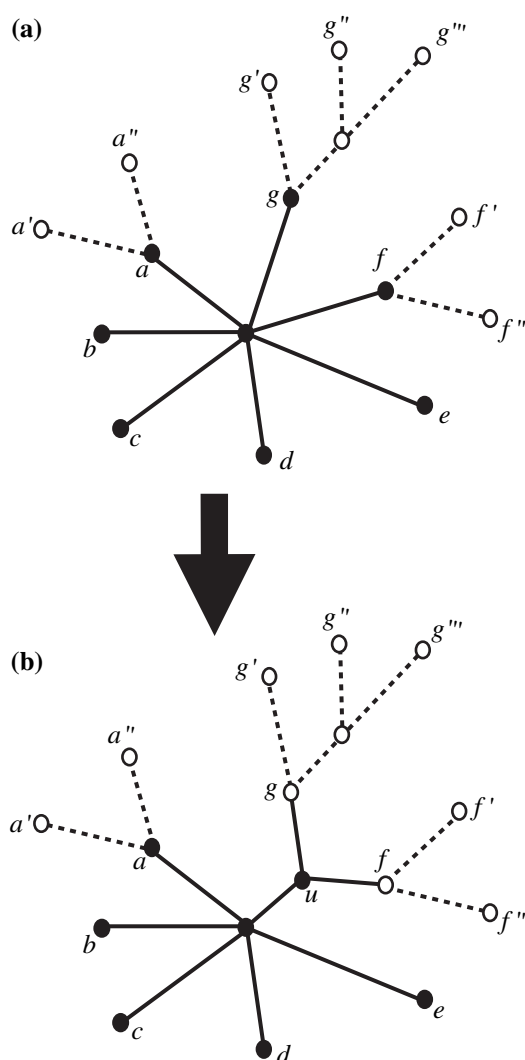
(a)

(b)

Fig. 1.—One NJ agglomeration step. In the current tree (*a*), the taxon set contains *a*, *b*, *c*, *d*, *e*, *f*, and *g*; some are original taxa, whereas the others (i.e., *a*, *f*, and *g*) correspond to subtrees built during the previous steps. Tree (*b*): after selection of the (*f*, *g*) pair, a new subtree is built, and both *f* and *g* are replaced by a unique taxon denoted as *u*. NJ terminates when the central node is fully resolved.

not to the subsequent steps; assuming that the current taxon set contains agglomerated nodes (e.g., *a*, *f*, and *g* in fig. 1), the OLS tree length estimate of the agglomerated tree (fig. 1*b*) differs from the estimate of Saitou and Nei. Thus, NJ performs local optimization to select taxon pairs but is not truly guided by minimization of the OLS tree length estimate. Several simulation studies (Saitou and Imanishi 1989; Kumar 1996; Gascuel 2000) showed that trees being shorter (in the OLS sense) than NJ trees are easily found, but these simulations also showed that these shorter trees are (slightly) less accurate than NJ trees, thus demonstrating that 1) ME + OLS does not fit the phylogenetic requirements well and 2) NJ cannot be viewed as optimizing this criterion.

Due to this confused situation (consistency, biological relevance, and the fact that NJ does not fully minimize the OLS tree length estimate), numerous authors (e.g., Swofford et al. 1996, p 490; Gascuel 2000; Felsenstein 2004, p 169) wrote that NJ does not explicitly optimize
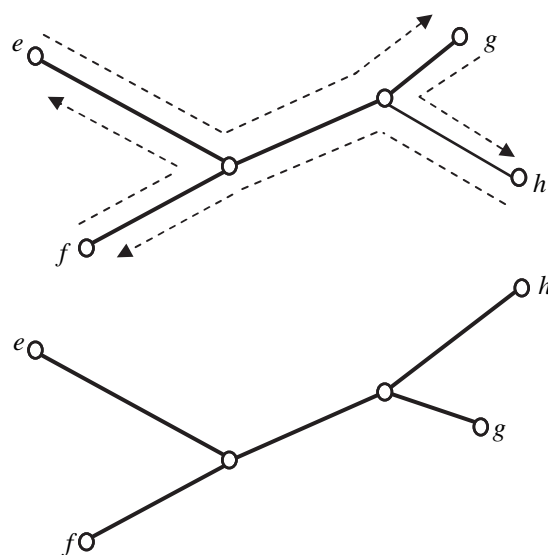

Fig. 2.—Circulars orders and tree length estimates. Two drawings of the same tree, which give rise to 2 different circular orders of the taxa and 2 estimates for *l* as a weighted sum of pairwise distances between leaves.

any criterion. In fact, NJ greedily optimizes a natural tree length estimate, as we shall explain now. In that respect, NJ does not differ from the other usual phylogenetic methods, for example, based on parsimony or maximum likelihood. It performs a heuristic search of tree space where each step is guided by this (global) tree length criterion. However, just as other standard methods, it does not explore the whole space and has no guarantee of finding the "best" tree for the criterion it seeks to optimize.

Given a tree with branch lengths, its length (*l*) is the total sum of all its branch lengths. We can express *l* as a weighted sum of pairwise distances between pairs of taxa, as has been observed by various authors (e.g., Yushmanov 1984; Makarenkov and Leclerc 1997; Sumiyama et al. 2001). For example, the reader should convince oneself that for the tree at the top of figure 2 we can write

$$l = \frac{1}{2}[d(e, g) + d(g, h) + d(h, f) + d(f, e)], \quad (4)$$

which corresponds to selecting pairs of leaves that traverse the tree in a clocklike fashion, that is, in the order *e*, *g*, *h*, *f*; because these 4 paths (shown dashed in the fig. 2) cover every edge twice, we divide by 1/2. There are other ways to express *l* as a weighted sum of the distance values, for example, if we redraw the "same" tree as shown in the bottom of figure 2 so that we traverse its leaves in the clocklike order *e*, *h*, *g*, *f*. This arbitrary dependence on how one chooses to draw the tree means that tree length estimates like equation (4) are somewhat unsatisfactory, and it raises the question of whether there is a "natural" way of writing *l* as a weighted sum of *d* values that does not depend on such choices. Pauplin (2000) found such a formula for binary trees, and this representation was subsequently extended to nonbinary trees by Semple and Steel (2004). This more general formula states

$$l = \sum_{\{i,j\}} w(i,j)d(i,j), \quad (5)$$

where the weight $w(i,j)$ is obtained as follows: consider the directed path from i to j, and for each interior node count how many outgoing branches are encountered—multiply these numbers together and divide 1 by the result, this gives $w(i,j)$. For example, with tree of figure $1a$, $w(a', a'') = 1/2$, $w(b, c) = 1/6$, and $w(a', g'') = 1/(2 \times 6 \times 2 \times 2) = 1/48$. For the tree in figure 2, this gives a different estimate of l to equation (4) because all 6 $w(i,j)$ values are nonzero and with some equal to 1/4 rather than 1/2. Semple and Steel (2004) showed that the estimate in equation (5) is precisely what one obtains by averaging all of the "simple" estimates (like eq. 4) over all possible ways to draw the tree in the plane. For binary trees, equation (5) simplifies to Pauplin's formula, which sets $w(i,j)$ equal to 1/2 raised to the power of the number of interior nodes on the path connecting i and j.

As Pauplin observed, this suggests a new version of the ME principle: simply select the tree that minimizes tree length estimate in equation (5), instead of using OLS estimator as proposed by Saitou and Nei. We designed fast algorithms following this new "balanced minimum evolution" (BME) scheme (Desper and Gascuel 2002), both to construct an initial tree by iteratively adding taxa on a growing tree and to refine this starting tree by nearest neighbor interchanges (Semple and Steel 2003, p 32). These algorithms are implemented in the FastME software (http://atgc.lirmm.fr/fastme). We observed with simulated data excellent topological accuracy of this new approach, that is, better than NJ and also better than other available distance methods. These findings were confirmed by Vinh and von Haeseler (2005) on very large phylogenies (up to 5,000 taxa). This suggested that the balanced scheme is well suited in phylogenetics, which we formally showed (Desper and Gascuel 2004) by demonstrating that equation (5) corresponds to a form of WLS tree length estimation, which puts more confidence on the short evolutionary distances than on the larger ones, for example, in our above example with figure $1a$, a and a' are neighboring and $d(a', a'')$ has weight 1/2, whereas a' and g'' are at 2 tree extremities and $d(a', g'')$ has weight 1/48. BME consistency was also shown in this paper.

Equation (5) also provides a way of quantifying any modification of an existing tree by considering the change in the l value. In particular, we can consider the selection of a pair of taxa, as occurs in the selection of neighbors by NJ (fig. 1), and this brings us to the main result of this note, whose proof appears in Desper and Gascuel (2005).

Theorem. The NJ method, as defined by equations (1), (2), and (3), selects at each step as neighbors that pair of current taxa, which most decreases the whole tree length, as computed using the generalized Pauplin formula (eq. 5).

Whole tree means that we consider all subtrees resulting from previous agglomerations and not only the central node. With figure 1, this implies that among all possible agglomerations within the current taxon set (i.e., $(a,b), (a,c), \ldots,$ until (f,g)), NJ selects that pair which minimizes the length of the whole tree shown in figure $1b$, including the subtrees and the original taxa a', a'', f', f'', etc. In other words, NJ greedily minimizes the BME score. Moreover, this explains why FastME performs better than NJ: it is

based on the same BME criterion but optimizes it further via topological rearrangements. This does not provide any guarantee on a given data set as the correct tree may not be the one with best fit, whatever the phylogenetic criterion being optimized (e.g., see Nei et al. 1998). But, on average, topological accuracy is improved by intensifying criterion optimization, as clearly shown by the comparisons of Vinh and von Haeseler (2005) between NJ and FastME.

We end by noting some other recent mathematical insights into NJ. Bryant (2005) has provided a different way of viewing the way that NJ selects pairs of taxa. Rather than showing that this selection optimizes some global quantity (the BME score) at each step, Bryant lists 3 desirable properties that any method for building trees from distances should possess if it operates (like NJ does) by successively grouping pairs of taxa. These properties are related to consistency and to taxon exchangeability. Assuming this, he proves that NJ selection criterion is unique among all possible linear criteria. In a further development, Levy et al. (2005) have shown how to modify the NJ algorithm to move beyond pairwise distances to more informative multiway distances (as might be obtained using maximum likelihood on sequences). Most recently, Mihaescu et al. (2006) have verified a conjecture of Atteson (1999) by establishing a robustness property of NJ to small perturbations in the data.

Acknowledgments

Literature Cited

Atteson K. 1999. The performance of the neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25:251–78.

Bryant D. 2005. On the uniqueness of the selection criterion in neighbor-joining. J Classif 22:3–15.

Desper R, Gascuel O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol 9:687–705.

Desper R, Gascuel O. 2004. Theoretical foundations of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. Mol Biol Evol 21:587–98.

Desper R, Gascuel O. 2005. The minimum evolution distance-based approach to phylogenetic inference. In: Gascuel O, editor. Mathematics of evolution & phylogeny. Oxford, UK: Oxford University Press. p 1–32.

Felsenstein J. 2004. Inferring phylogenies. Sunderland, MA: Sinauer Associates.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. Science 155:279–84.

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. Mol Biol Evol 11: 961–3.

Gascuel O. 1997. Concerning the NJ algorithm and its unweighted version, UNJ. In: Mirkin B, McMorris FR, Roberts FS, Rzhetsky A, editors. Mathematical hierarchies and biology. DIMACS series in discrete mathematics and theoretical computer science. Providence, RI: American Mathematical Society. p 149–70.

Gascuel O. 2000. On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. Mol Biol Evol 17:401–5.

Kimura M. 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J Mol Evol 16:111–20.

Kuhner MK, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–68.

Kumar S. 1996. A stepwise algorithm for finding minimum evolution trees. Mol Biol Evol 13:584–93.

Levy D, Yoshida R, Pachter L. 2005. Beyond pairwise distances: neighbor joining with phylogenetic diversity estimates. Mol Biol Evol (Advanced access, November 9, 2005).

Makarenkov V, Leclerc B. 1997. Circular orders of tree metrics, and their uses for the reconstruction and fitting of phylogenetic trees. In: Mirkin B, McMorris FR, Roberts F, Rzhetsky A, editors. Mathematical hierarchies and biology. DIMACS series in discrete mathematics and theoretical computer science. Providence, RI: American Mathematical Society. p 183–208.

Mihaescu R, Levy D, Pachter L. 2006. Why neighbour-joining works. arXiv:cs.DS/0602041 v1.

Mirkin B. 1996. Mathematical classification and clustering. London: Kluwer Academic Publishers.

Nei M, Kumar S. 2000. Molecular evolution and phylogenetics. Oxford, UK: Oxford University Press.

Nei M, Kumar S, Takahashi K. 1998. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc Natl Acad Sci 95:12390–7.

Pauplin Y. 2000. Direct calculation of a tree length using a distance matrix. J Mol Evol 51:41–7.

Rzhetsky A, Nei M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol Biol Evol 10:1073–95.

Saitou N. 1996. Reconstruction of gene trees from sequence data. In: Doolittle R, editor. Methods in enzymology. Volume 266. Orlando, FL: Academic Press, Inc. p 427–49.

Saitou N, Imanishi M. 1989. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic reconstructions in obtaining the correct tree. Mol Biol Evol 6:514–25.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstruction of phylogenetic trees. Mol Biol Evol 4:406–25.

Semple C, Steel M. 2003. Phylogenetics. Oxford, UK: Oxford University Press.

Semple C, Steel M. 2004. Cyclic permutations and evolutionary trees. Adv Appl Math 32:669–80.

Studier JA, Keppler KJ. 1998. A note on the neighbor-joining method of Saitou and Nei. Mol Biol Evol 5:729–31.

Sumiyama K, Kim CB, Ruddle FH. 2001. An efficient cis-element discovery method using multiple sequence comparisons based on evolutionary relationships. Genomics 71:260–2.

Swofford DL, Olsen GL, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, editors. Molecular sytematics. Sunderland, MA: Sinauer. p 407–514.

Vinh LS, von Haeseler A. 2005. Shortest triplet clustering: reconstructing large phylogenies using representative sets. BMC Bioinformatics 6:92.

Yushmanov SV. 1984. Construction of a tree with p leaves from $2p$-3 elements of its distance matrix (Russian). Matematicheskie Zametki 35:877–87.