Full Details of the Weighbor Method

William J. Bruno, Nicholas D. Socci, and Aaron L. Halpern

Here we present all of the formulas used in the Weighbor program as of September, 1999. We chose formulas that we thought would give the best estimates in a single step. In some cases simpler formulas might give equally good results, or iterative formulas might give some improvement.

The scoring function $S(i,j)$

The central task is to compute $S(i,j)$. We can write it as

$$S(i,j) \equiv a_{ij}(\Delta^2 b_{ij} - (\Delta b_{ij})^2)g - 2\ln\left[\frac{1}{2}\mathrm{erfc}\left(\frac{-z_{PQ}}{\sqrt{2}}\right)\right] \tag{1}$$

All quantities are defined below except g, which is discussed in the paper, and equals $1/(N-3)$. This S is twice the S in the paper. The sums a_{ij}, Δb_{ij} and $\Delta^2 b_{ij}$ are stored and updated from one iteration to the next:

$$\Delta b_{ij} \equiv \frac{1}{a_{ij}} \sum_{k\notin\{i,j\}} \frac{1}{\sigma^2_{ik;j} + \sigma^2_{jk;i}}(d_{ik} - d_{jk}) \tag{2}$$

$$a_{ij} \equiv \sum_{k\notin\{i,j\}} \frac{1}{\sigma^2_{ik;j} + \sigma^2_{jk;i}} \tag{3}$$

$$\Delta^2 b_{ij} \equiv \frac{1}{a_{ij}} \sum_{k\notin\{i,j\}} \frac{1}{\sigma^2_{ik;j} + \sigma^2_{jk;i}}(d_{ik} - d_{jk})^2 \tag{4}$$

If i and j are neighbors, then Δb_{ij} is an estimate of $d_{iP} - d_{jP}$, where P joins i and j. If the tree is a star phylogeny, then it can be shown that Δb_{ij} is in fact the generalized least-squares estimate, provided $\sigma^2_{ik;j} \equiv \sigma^2_{\text{non-add.}}(d_{iP}, d_{Pk})$ equals the non-additive variance.

$$\sigma^2_{ik;j} = \sigma^2_{\text{non-add.}}(d_{iP}+c_i, d_{Pk}+c_k) \equiv {\sigma_{\mathrm{t}}}^2(d_{ik}+c_i+c_k) - {\sigma_{\mathrm{t}}}^2(d_{iP}+c_i) - {\sigma_{\mathrm{t}}}^2(d_{Pk}+c_k) \tag{5}$$

The c_i are computed below. We compute d_{iP} assuming additivity:

$$d_{iP} = \max(0, \min(d_{ik}, (d_{ij} + d_{ik} - d_{jk})/2)) \tag{6}$$

with the min and max added to deal with cases where the triangle inequality is violated. Note that this estimate implies $d_{ik} = d_{iP} + d_{Pk}$, which seems to imply zero non-additivity. This just confirms that our d_{iP} estimate ignores non-additivity, but this is the best estimate available without knowing the tree.

We define $z_{PQ} \equiv d_{PQ}/\sigma_{PQ}$ and estimate it by

$$z_{PQ} = \min_{k,l \notin \{i,j\}} \frac{d_{PQ}(ij,kl)}{\sqrt{\sigma^2_{PQ}(ij,kl) + (\sigma^2_{ij;k} + \sigma^2_{ij;l})/8}} \tag{7}$$

where the min over k and l is approximated below for increased speed, and we use the estimates

$$d_{PQ}(ij,kl) = [\frac{\frac{d_{ik}+d_{jl}}{\sigma^2_{ik;\min}+\sigma^2_{jl;\min}} + \frac{d_{il}+d_{jk}}{\sigma^2_{il;\min}+\sigma^2_{jk;\min}}}{\frac{1}{\sigma^2_{ik;\min}+\sigma^2_{jl;\min}} + \frac{1}{\sigma^2_{il;\min}+\sigma^2_{jk;\min}}} - d_{ij} - d_{kl}]/2 \tag{8}$$

$$\sigma^2_{PQ}(ij,kl) = [\frac{1}{\frac{1}{\sigma^2_{ik;\min}+\sigma^2_{jl;\min}} + \frac{1}{\sigma^2_{il;\min}+\sigma^2_{jk;\min}}} + (\sigma^2_{kl;i} + \sigma^2_{kl;j})/2]/4 \tag{9}$$

where $\sigma^2_{ik;\min}$ stands for $\min(\sigma^2_{ik;j}, \sigma^2_{ik;l})$; the reason this min appears is described in the paper. The essence of these formulas is that there are two partly independent ways to compute d_{PQ}, and we take their weighted average.

The $\min_{kl}$ in equation 7 implies order N^2 steps, which is incompatible with our goal of an N^3 method overall. Instead we employ a "greedy" heuristic to find k and l that nearly minimize z_{PQ}. We start with a trial value of $k = q_2[i]$, which is the taxon that placed second in the tournament (see below) that determined $q[i]$. Then, holding k fixed, we try all values $l \notin \{i,j,k\}$ to find the l that minimizes z_{PQ}. Taking this value of l as fixed, we now minimize over $k \notin \{i,j,l\}$.

The value of z_{PQ} for this k and l is returned. We expect this heuristic to work well for nearly additive distances; it can be proven that for additive distances it will return a negative z_{PQ} if the true minimum is negative.

The tournament for $q[i]$ using $\Delta S(i,j,j')$

The pair to be joined could be found by evaluating $S(i,j)$ for all pairs (i,j), but this would still imply an N^4 algorithm. Instead we take advantage of a constant time approximation for z_{PQ} that is applicable to evaluating the relative likelihood of two pairs that share a taxon in common, i.e., evaluating $S(i,j)-S(i,j')$. This allows us to do a set of single elimination tournaments to find the best taxon to join to each taxon i. This means, for each i arbitrarily taking the first $j \neq i$ to be the current winner, and competing j against the next candidate j'. The winner will then go on to compete against the next candidate, and so on. After each taxon has had a chance to compete, the winner is declared and stored as $q[i]$. The contest between j and j' is determined by the sign of the quantity $\Delta S(i,j,j')$:

$$\Delta S(i,j,j') \equiv$$

$$[a_{ij}(\Delta^2 b_{ij} - \Delta b_{ij}^2) - a_{ij'}(\Delta^2 b_{ij'} - \Delta b_{ij'}^2)]h - 2\ln\left[\frac{\mathrm{erfc}\left(\frac{d_{ij} - \overline{\Delta b''_{jj'}} - \overline{d_{ij'}}}{\sqrt{2(\sigma^2_{ij;j'} + \overline{1/a''_{jj'} + \sigma^2_{ij';j}})}}\right)}{\mathrm{erfc}\left(\frac{d_{ij'} - \overline{\Delta b''_{j'j}} - \overline{d_{ij}}}{\sqrt{2(\sigma^2_{ij';j} + \overline{1/a''_{j'j} + \sigma^2_{ij;j'}}))}}\right)}\right] \tag{10}$$

The meaning of the double primes ($''$) and overbars ($\overline{}$) are defined below. The factor h is analogous to g above, and in fact the values we use are the same. The quantity inside the erfc() is similar to what we had in $S(i,j)$; the difference is that we assume that the node Q is defined by the common ancestor between j' and the $N-3$ other taxa; i.e., we compute z_{PQ} holding $k = j'$, and averaging over l.

The $''$ means any term involving the triplet i, j, j' is excluded (subtracted off) from the sum. Thus

$$\Delta b''_{jj'} \equiv [a_{jj'} \Delta b_{jj'} - (d_{ji} - d_{j'i})/(\sigma^2_{ji;j'} + \sigma^2_{j'i;j})]/a''_{jj'} \tag{11}$$

and

$$a''_{jj'} \equiv a_{jj'} - 1/(\sigma^2_{ji;j'} + \sigma^2_{j'i;j}) \tag{12}$$

These terms are subtracted because we want to average over the fourth taxon l subject to $l \notin \{i, j, j'\}$.

The overlines above the various quantities could be ignored, but a more efficient estimate is obtained by using a weighted average of the two distinct ways to estimate d_{PQ} (namely, interchanging l and j'), achieved using:

$$\overline{d_{ij'}} \equiv \frac{d_{ij'}/(\sigma^2_{ij';j} + 1/a''_{jj'}) + d_{jj'}/(\sigma^2_{jj';i} + 1/a''_{ij'})}{1/(\sigma^2_{ij';j} + 1/a''_{jj'}) + 1/(\sigma^2_{jj';i} + 1/a''_{ij'})} \tag{13}$$

$$\overline{\Delta b''_{jj'}} \equiv \frac{\Delta b''_{jj'}/(\sigma^2_{ij';j} + 1/a''_{jj'}) + \Delta b''_{ij'}/(\sigma^2_{jj';i} + 1/a''_{ij'})}{1/(\sigma^2_{ij';j} + 1/a''_{jj'}) + 1/(\sigma^2_{jj';i} + 1/a''_{ij'})} \tag{14}$$

and

$$\overline{1/s''_{jj'} + \sigma^2_{ij';j}} \equiv \frac{1}{1/(\sigma^2_{ij';j} + 1/s''_{jj'}) + 1/(\sigma^2_{jj';i} + 1/s''_{ij'})} \tag{15}$$

The winners of the tournaments thus define $q[i]$ for all i. We restrict our consideration to the pairs $(i, q[i])$ that satisfy $q[q[i]] = i$ (unless none do); this requirement—that both taxa won the tournaments to join to the other—is discussed later. We proceed to join the pair in this restricted set that minimizes $S(i, q[i])$.

Creating the New Node

Once we have decided which pair to join, we must prepare for the next iteration. Suppose we have decided to join taxon i with taxon i'. Calling the newly created node P, the distance from i to P is computed as

$$d_{iP} \equiv \max(0, \min(d_{ii'}, (\Delta b_{ii'} + d_{ii'})/2)) \tag{16}$$

The min and max only come into play when the distances violate the triangle inequality. In cases where $z(i,i') < 0$, a correction to this formula is applied, as discussed below.

The distance from P to some other taxon $j \notin \{i, i'\}$ is given by the weighted average

$$d_{P\,j} = \frac{(d_{ij} - d_{iP})/\sigma^2_{\text{avg}}(iP) + (d_{i'j} - d_{i'P})/\sigma^2_{\text{avg}}(i'P)}{1/\sigma^2_{\text{avg}}(iP) + 1/\sigma^2_{\text{avg}}(i'P)} \tag{17}$$

where $\sigma^2_{\text{avg}}(iP)$ is the mean squared non-additivity $\sigma^2_{ij;i'}$ averaged over j. This method of averaging was put forward by Gascuel (1997) for BIONJ. It is important that the same weights are used for all j because this causes errors in $d_{ii'}$ to act as additive errors in later iterations.

The expected error in $d_{P\,j}$ will be greater (or equal if $d_{iP} = 0$ or $d_{i'P} = 0$) than the error expected for an actual sequence at P. We keep track of this increased error by introducing a quantity c_P which an amount of distance to add to d_{Pj} so that the expected error comes out about right. Again, having c_P independent of j helps keep local errors additive in later iterations. The c's require us to change the non-additive variance formula to

$$\sigma^2_{ij;k} = \sigma^2_{\text{non-add.}}(d_{iP}+c_i, d_{jP}+c_j) \equiv {\sigma_{\mathrm{t}}}^2(d_{ij}+c_i+c_j) - {\sigma_{\mathrm{t}}}^2(d_{iP}+c_i) - {\sigma_{\mathrm{t}}}^2(d_{jP}+c_j) \tag{18}$$

The c's are zero when i and j are actual sequences. To compute the c_P that will give the appropriate non-additive variance, we need the inverse of the σ_t^2 function, which we write $[\sigma_t^2]^{-1}$. For JC,

$$[\sigma_t^2]^{-1}(x) = \frac{\beta - 1}{\beta} \ln\left(\frac{2[x\beta^2 L + (\beta-1)^2]}{\beta\sqrt{4x(\beta-1)L + (\beta-1)^2} + (\beta-1)(\beta-2)}\right) \tag{19}$$

where β is the alphabet size. If more complex models are implemented, this function could be computed numerically. We obtain c_P by plugging into $[\sigma_t^2]^{-1}$ the amount of extra non-additive variance expected in distances computed to P compared to what would be expected if there were a leaf at P. This is computed by propagating the variances in the branch lengths d_{iP} and $d_{i'P}$, plus any c's associated with them, through the weights in equation 17. The variance in $d_{ii'}$ also contributes to errors in d_{Pj}, but this contribution is the same for all j, and is therefore ignored because it does not contribute to non-additivity in subsequent iterations. We obtain

$$c_P = [\sigma_t^2]^{-1}\left(\frac{\frac{\sigma_t^2(c_i+b_{i;i'})}{(\sigma^2_{\text{avg}}(iP))^2} + \frac{\sigma_t^2(c_{i'}+b_{i';i})}{(\sigma^2_{\text{avg}}(i'P))^2}}{\left(\frac{1}{\sigma^2_{\text{avg}}(iP)} + \frac{1}{\sigma^2_{\text{avg}}(i'P)}\right)^2}\right) \tag{20}$$

Correcting for negative z

If z is negative for the pair being joined, then just using additivity to estimate d_{iP} will result in a negative d_{PQ}, so we want to invoke the positivity constraint as well. A negative $z(i,i')$ suggests a positive error in $d_{ii'}$, but errors in other distances could also be the cause. We seek the most likely compromise distances subject to $d_{PQ} \geq 0$. First, we compute a better, order N^2 estimate of z defined by:

$$z(ij) \equiv d_{PQ}(i\,j)/\sqrt{\sigma^2_{PQ}(ij)} \tag{21}$$

where

$$d_{PQ}(ij) \equiv \frac{1}{a_{PQ}(ij)} \sum_k \frac{1}{\sigma^2_{PQ}(ij,kl)} \min_l [d_{PQ}(ij,kl)] \tag{22}$$

in which $a_{PQ}(ij)$ is the normalization of this weighted average (namely, the sum over k of the $1/\sigma^2_{PQ}(ij,kl)$ for the l that minimizes d_{PQ}) and

$$\sigma^2_{PQ}(ij) \equiv \frac{1}{a_{PQ}(ij)} + \frac{1}{4(N-2)} \sum_{k \notin \{i,j\}} \sigma^2_{ij;k} \tag{23}$$

Then define

$$\phi = \frac{-d_{PQ}(ii')a_{PQ}(ii')/2}{1/\sigma^2_{ii'} + a_{PQ}(ii')/4} \tag{24}$$

where $\sigma^2_{ii'}$ is the average of $\sigma^2_{ii';k}$ over k. This ϕ estimates the amount by which $d_{ii'}$ was most likely overestimated, according to our variance model. Once ϕ is determined, if it is positive (which it is whenever the full z-score is negative), it is subtracted from $d_{ii'}$ for all of the calculations involving P. Thus d_{iP} and $d_{i'P}$ are reduced, and this causes distances from P to all the other taxa to increase. When the node Q is eventually created, d_{PQ} will usually still come out negative, but by a smaller amount. Actually, it will be forced to zero by the triangle inequality in any case. Thus the main effect of these efforts is that d_{iP} and $d_{i'P}$ are reduced to more likely values and d_{Qj} will be increased for all $j \notin \{i, i'\}$.

Note, in the paper, j appears in place of i' in the above formulas; h is used in place of ϕ, (h is not mentioned in the paper).

An iteration of the method is completed by replacing nodes i and i' by P, and reducing N by 1. When only three taxa remain, their distances are computed assuming additivity (as in NJ), except that we rely on the shorter distances in the case of a triangle inequality violation. We should also note that to avoid problems caused by zero denominators, such as when an estimated variance is zero, we add a small quantity ϵ/L, where $\epsilon = 10^{-9}$, to most denominators.

The $q[q[i]] = i$ requirement

With exact distances all true neighbors would satisfy $q[q[i]] = i$. Because the $q[q[i]] = i$ requirement may also be satisfied by some (but not all) incorrect pairs when distances are exact, this requirement will cause some bias in the types of errors the method makes.

For example, the tree in Figure 4 of the paper is neutral with respect to biases except when the length b of the connecting branch is very short, and in this case Weighbor will tend to incorrectly join the taxa closest to this short branch. We find this to be a very slight effect: the disadvantageous bias of the $q[q[i]] = i$ requirement results in a reduction in performance of only about 4 trees per 1000 when $b = 0.01$ in Figure 4; Weighbor still significantly outperforms NJ and BIONJ. Conversely, if the tree is changed so that these close branches are neighbors, the requirement improves performance by about 27 trees per 1000, owing to the improvement in sensitivity. The program does allow the requirement to be turned off, using the "-Xq" option, if one decides to sacrifice some sensitivity for the sake of being as unbiased as possible.