# Appendix

*Solution of Simultaneous Equations in the Meta-NJ Algorithm*

The simultaneous equations $(1)$ and $(2)$ that define the tree topologies associated to internal vertices in the meta-NJ algorithm are solved in the following way. Consider each split $p$ separately: equations $(1)$ and $(2)$ become

$$p \in X \Leftrightarrow I_A(p) + I_B(p) + I_Z(p) > 1,$$
$$p \in Z \Leftrightarrow I_X(p) + \sum_{i=1}^{k} I_{Z_i}(p) > \lfloor \frac{k+1}{2} \rfloor$$

where $I_R(p)$ is the indicator function for $p \in R$ for any set of splits $R = A, B, Z_i$. The square brackets $\lfloor x \rfloor$ denote the largest integer less than or equal to $x$. There are four possible configurations for membership of $X$ and $Z$:

$$p \in X, p \in Z \quad \Rightarrow \quad I_A(p) + I_B(p) > 0, \quad \text{and} \quad \sum I_{Z_i}(p) > \lfloor \frac{k-1}{2} \rfloor \qquad (C1)$$
$$p \in X, p \notin Z \quad \Rightarrow \quad I_A(p) = I_B(p) = 1, \quad \text{and} \quad \sum I_{Z_i}(p) \leq \lfloor \frac{k-1}{2} \rfloor \qquad (C2)$$
$$p \notin X, p \in Z \quad \Rightarrow \quad I_A(p) = I_B(p) = 0, \quad \text{and} \quad \sum I_{Z_i}(p) > \lfloor \frac{k+1}{2} \rfloor \qquad (C3)$$
$$p \notin X, p \notin Z \quad \Rightarrow \quad I_A(p) + I_B(p) \leq 1, \quad \text{and} \quad \sum I_{Z_i}(p) \leq \lfloor \frac{k+1}{2} \rfloor \qquad (C4)$$

If the conditions $C1$–$C4$ on the right were mutually exclusive, they would then be necessary and sufficient to determine whether $p$ should be included in $X$ and / or $Z$. This is very nearly the case: only conditions $C1$ and $C4$ can be satisfied simultaneously, when

$$p \in A \triangle B, \text{ and } \sum_{1}^{k} I_{Z_i}(p) = \lfloor \frac{k+1}{2} \rfloor$$

where $A \triangle B$ denotes the symmetric difference of $A$ and $B$. Under such circumstances it is necessary to decide whether to include $p$ in both $X$ and $Z$, or whether to exclude $p$ from both sets. Figure A-1 illustrates the situation in the two cases $k$ even and $k$ odd, where $k$ is the number of leaves attached to $Z$. When $k$ is odd, the score of the tree is minimised by including $p$ in both $X$ and $Z$. However, when $k$ is even the total score does not change whether $p$ is included or excluded from $X$ and $Z$. By including $p$, the two sets $X$ and $Z$ may fail to be tree-like, so we exclude it. This is similar to the construction of a median tree for an even number of tree topologies, when splits that are in exactly half the topologies are excluded from the majority consensus.

The complete set of rules for solving equations $(1)$ and $(2)$ is therefore as follows. For each split $p$ first determine which conditions $C1$–$C4$ are met. If only one condition is satisfied, assign $p$ to $X$ and $Z$ according to the left-hand side of the equations above. In the unique case that both $C1$ and $C4$ are satisfied, $p$ is included in both $X$ and $Z$ when $k$ is odd, but excluded from both sets when $k$ is even.

To check that the solutions $X$ and $Z$ are tree-like, it is sufficient to show this only for $X$, since then $Z$ is defined as the majority consensus of tree-like sets of splits and so must be tree-like itself. Suppose there are two incompatible splits $p, q$ in $X$. Then, since $X \subset A \cup B$ (conditions $C1$ and $C2$), it follows that one split must be from $A \setminus B$ and the other from $B \setminus A$. Both splits must be in two neighbours of $X$ so both are elements of $Z$. The rules specified above then guarantee that $p$ and $q$ are in the majority consensus of $T_{Z_1}, \ldots, T_{Z_k}$ and are therefore compatible, so we have a contradiction. It follows that the meta-NJ algorithm is well-defined and builds tree-like sets of splits at each step.

*Satisfying the Local Optimality Criterion*

Suppose at some stage in the meta-NJ algorithm the tree $\hat{T}_r$ satisfies the local optimality criterion at each vertex (namely that each tree $T_{\hat{v}}$ is the majority consensus of the neighbours of $\hat{v}$). Consider a refinement $\hat{T}_r^{AB}$ formed by the agglomeration of two vertices $A, B$. It is shown below that, following the agglomeration:

1. $T_A$ and $T_B$ always satisfy the local optimality criterion, and
2. $T_{Z_1}, \ldots, T_{Z_k}$ satisfy the criterion provided certain exceptional splits are removed.

This last point leads to a slight modification of the agglomerative step of the meta-NJ algorithm in which some splits are pruned from the meta-tree in order to ensure local optimality.

First consider vertices $A$ and $B$. Suppose that $A$ is the agglomeration of two nodes $A_1$ and $A_2$ (that might themselves have descendants) as shown in Figure A-2. Also suppose that every node in $\hat{T}_r$ satisfies the local optimality criterion so that

$$T_A = \text{maj}\{T_{A_1}, T_{A_2}, T_Y\} \text{ and} \tag{A-1}$$

$$T_Y = \text{maj}\{T_A, T_B, T_{Z_1}, \ldots, T_{Z_k}\}. \tag{A-2}$$

We want to show that $T_A = M$ where $M$ is the majority consensus of $T_{A_1}, T_{A_2}, T_X$, so that after agglomeration vertex $A$ is still locally optimal. Fix a split $p \in T_A$. If $p \in T_{A_1} \cap T_{A_2}$ then $p$ is also trivially contained in $M$, so suppose that $p$ is in just one of $T_{A_1}$ and $T_{A_2}$. It follows from (A-1) that $p \in T_Y$. Using the fact that $p$ is contained in the right-hand side of equation (A-2), the rules for determining $T_X$ and $T_Z$ defined above can be followed, and they show that $p \in T_X$. It follows that $p \in M$, and this establishes that $T_A \subset M$. Conversely, pick any $p \in M$ and without loss of generality assume that $p$ is contained in just one of $T_{A_1}$ and $T_{A_2}$. Since $p$ is necessarily in $T_X$ either $C1$ holds but not $C4$, or both $C1$ and $C4$ hold with $k$ odd. Both these conditions imply that $p \in \text{maj}\{T_A, T_B, T_{Z_1}, \ldots, T_{Z_k}\} = T_Y$, and so $T_A = M$. By symmetry the local optimality condition also holds for vertex $B$.

Next consider applying similar reasoning to the vertices $Z_1, \ldots, Z_k$. In particular suppose that $Z_1$ is the agglomeration of two nodes $U, V$. The argument above can be repeated for a split $p \in T_{Z_1}$, the aim being to prove that if $p \in T_{Z_1} = \text{maj}\{T_U, T_V, T_Y\}$ then $p$ is in $\text{maj}\{T_U, T_V, T_Z\}$ after the agglomerative step. The non-trivial case occurs when $p \in Y$ but $p$ is only in one of $T_U, T_V$. We want to show that the rules for determining $T_Z$ ensure that it contains $p$. This is nearly always the case but there is one exception: when $k$ is odd, $p \in T_A \cap T_B$, and $\sum I_{Z_i}(p) = \frac{k-1}{2}$, condition $C2$ holds and

$p \notin T_Z$. In this situation, illustrated by Figure A-3, $T_{Z_1}$ is no longer the majority consensus of its neighbours after agglomeration. However, this can be fixed by removing the split $p$ from $T_{Z_1}$ (and from the descendants of vertex $Z_1$ if necessary). The meta-NJ algorithm can therefore be extended via the following step in order to ensure local optimality:

3(b) When $k$ is odd, identify contained in $T_A$ and $T_B$, and which satisfy $\sum I_{Z_i}(p) = \frac{k-1}{2}$. Remove these splits from the subtrees hanging from vertices $Z_1, \ldots, Z_k$ so that the local optimality condition is satisfied.

In practice this step seems to be required rarely; for example, it was not required for any of the meta-trees constructed in the Applications section.

Figure A-1: Meta-trees and a split $p$ satisfying both conditions $C1$ and $C4$. Each vertex is labelled $p$ if it contains the split; $\not p$ indicates that $p$ is not contained in the vertex. (a) Odd number $k$ of leaves of $Z$: the score is minimised by including $p$ in $X$ and $Z$. (b) Even number $k$ of leaves of $Z$: $p$ is excluded from $X$ and $Z$, although the score is unaffected.

Figure A-2: One step of the meta-NJ algorithm when vertex $A$ has two descendants.

Figure A-3: One step of the meta-NJ algorithm when vertex $Z_1$ has two descendants. The split $p$ is contained in $T_{Z_1}$, but after agglomeration it is not contained in the majority consensus of the neighbours of $Z_1$.