Tree Metrics

8° Now that we have described a method of determining a mathematical relationship between two sequences, we wish to demonstrate how these relationships can be turned into a useful and simple plot: *the phylogenetic tree.*

> **Definition (phylogenetic tree).** A phylogenetic tree is a tree T together with a labeling of its leaves. It describes the evolutionary relationships between species; some formulations of phylogenetic trees (such as ours) describe the amount of time species took to evolve. The number of combinatorial types of phylogenetic trees with the same leaves grows exponentially. In phylogenetics, a typical problem is to select a tree, based on data, from the large number of possible choices.

Let us assume that we have four sequences of DNA from four separate but similar species of crab, and let us call them c_1, c_2, c_3, and c_4. Let us assume that we have aligned each sequence with each other and scored each alignment with a biologically appropriate scoring schema. We will call the resulting scores 'evolutionary distances'.

9° *The dissimilarity matrix.* Let us organize our evolutionary distances into a matrix. Let each row and column correspond to a sequence, and let each element correspond to the distance between each sequence. Then we have the following:

$$\begin{array}{cc} \epsilon & \begin{array}{cccc} c_1 & c_2 & c_3 & c_4 \end{array} \\ \begin{array}{c} c_1 \\ c_2 \\ c_3 \\ c_4 \end{array} & \begin{pmatrix} d(c_1,c_1) & d(c_2,c_1) & d(c_3,c_1) & d(c_4,c_1) \\ d(c_1,c_2) & d(c_2,c_2) & d(c_3,c_2) & d(c_4,c_2) \\ d(c_1,c_3) & d(c_2,c_3) & d(c_3,c_3) & d(c_4,c_3) \\ d(c_1,c_4) & d(c_2,c_4) & d(c_3,c_4) & d(c_4,c_4) \end{pmatrix} \end{array}.$$

Of course, using an appropriate scoring schema, we would hope that the distance derived from aligning a sequence with itself would be 0. Let us assume that we have determined all of the distances. Then we have the following *dissimilarity matrix D*:

$$D = \begin{array}{cc} \epsilon & \begin{array}{cccc} c_1 & c_2 & c_3 & c_4 \end{array} \\ \begin{array}{c} c_1 \\ c_2 \\ c_3 \\ c_4 \end{array} & \begin{pmatrix} 0 & 1.1 & 1.0 & 1.4 \\ 1.1 & 0 & 0.3 & 1.3 \\ 1.0 & 0.3 & 0 & 1.2 \\ 1.4 & 1.3 & 1.2 & 0 \end{pmatrix} \end{array}.$$

This matrix is easy to read: for example, the distance between c_2 and c_4 is 1.3.

10° If all is well, then it is simple to build a phylogenetic tree using the above distances. Unfortunately, a tree can only be built from a metric that fulfills certain properties. If our matrix does not fulfill these properties, then it is not a tree metric. Luckily, it is very easy to determine whether or not a dissimilarity matrix fits the properties of a tree metric using *the Four-Point Condition.*

> **Theorem 1 (The Four-Point Condition).** A metric d is a tree metric if and only if, for any four leaves u, v, x, y the maximum of the three numbers $d(u, v) + d(x, y)$, $d(u, x) + d(v, y)$ and $d(u, y) + d(v, x)$ is attained at least twice.

Let us find out if D satisfies the Four-Point Condition. We have only four sequences so our choices for u, v, x, y is obvious: $u = c_1, v = c_2, x = c_3$, and $y = c_4$. Let us calculate our three numbers:

$$\begin{aligned}
d(u, v) + d(x, y) &= (1.1) + (1.2) = 2.3 \\
d(u, x) + d(v, y) &= (1.0) + (1.3) = 2.3 \\
d(u, y) + d(v, x) &= (1.4) + (0.3) = 1.7.
\end{aligned}$$

Indeed, the max of the three numbers, 2.3, was achieved twice and hence D satisfies the Four-Point Condition. This means that D corresponds to a metric d_T for some phylogenetic tree T. In a case with more than four sequences, all combinations must be considered.

Let d_T be the following metric corresponding to five leaves a, b, c, d, and e:

$$\begin{array}{c c}
 & \begin{array}{ccccc} a & b & c & d & e \end{array} \\
\begin{array}{c} a \\ b \\ c \\ d \\ e \end{array} & \begin{pmatrix} 0 & 5 & 6 & 6 & 7 \\ 5 & 0 & 3 & 5 & 6 \\ 6 & 3 & 0 & 6 & 7 \\ 6 & 5 & 6 & 0 & 4 \\ 7 & 6 & 7 & 3 & 0 \end{pmatrix}
\end{array}.$$

Starting with five leaves, we have $\binom{5}{4} = 5$ distinct combinations of four leaves:

$$\begin{aligned}
c_1 &= \{a, b, c, d\} \\
c_2 &= \{e, b, c, d\} \\
c_3 &= \{a, e, c, d\} \\
c_4 &= \{a, b, e, d\} \\
c_5 &= \{a, b, c, e\}.
\end{aligned}$$

In order for d_T to be a tree metric, the conditions set forth in the theorem above must be satisfied for every combination of four leaves. Let us call the first number in the theorem A, and similarly B and C for the second and third. Starting with c_1,

$$\begin{aligned}
A_1 &= d(a,b)+d(c,d)=11\\
B_1 &= d(a,c)+d(b,d)=11\\
C_1 &= d(a,d)+d(b,c)=9\\
A_2 &= d(e,b)+d(c,d)=12\\
B_2 &= d(e,c)+d(b,d)=12\\
C_2 &= d(e,d)+d(b,c)=6\\
A_3 &= d(a,e)+d(c,d)=13\\
B_3 &= d(a,c)+d(e,d)=10\\
C_3 &= d(a,d)+d(e,c)=13\\
A_4 &= d(a,b)+d(e,d)=9\\
B_4 &= d(a,e)+d(b,d)=12\\
C_4 &= d(a,d)+d(b,e)=12\\
A_5 &= d(a,b)+d(c,e)=12\\
B_5 &= d(a,c)+d(b,e)=12\\
C_5 &= d(a,e)+d(b,c)=10.
\end{aligned}$$

Obviously, for each combination, the maximum of A_n, B_n, C_n is achieved at least twice. Hence our metric d_T is a tree metric.