1 Four point condition

In general, if a distance matrix is to be represented faithfully by a tree, it must satisfy the following four-point condition [1]

$$\mathcal{D}(i,j) + \mathcal{D}(m,n) \leq \max\{\mathcal{D}(i,m) + \mathcal{D}(j,n), \mathcal{D}(j,m) + \mathcal{D}(i,n)\}. \qquad (1.1)$$

which generalizes the familiar triangle inequality (take $m = n$).

This requirement implies that typical biological trees will not uniquely represent a given biological distance matrix.

Definition 1.1 *A matrix that satisfies the four point condition is called* additive.

Theorem 1.1 *A distance matrix can be represented by a tree if and only if it is additive.*

See handout for a proof of 'if' and a corresponding algorithm.

1.1 Four point condition: meaning

The four point condition

$$\mathcal{D}(i,j) + \mathcal{D}(m,n) \leq \max\{\mathcal{D}(i,m) + \mathcal{D}(j,n), \mathcal{D}(j,m) + \mathcal{D}(i,n)\}$$

encodes a special requirement among the three values in the inequality.

To see what the required relationship is, let us simplify notation. Suppose that we have three positive numbers, a_1, a_2 and a_3, and we require that

$$a_i \leq \max\{a_j, a_k\} \tag{1.2}$$

for any partition $\{i,j,k\} = \{1,2,3\}$.

Now relabel the numbers so that $a_1 \geq a_2 \geq a_3$.

Then (1.2) is equivalent to the statement $a_1 = a_2$.

That is, (1.2) is equivalent to the requirement that the largest two of the a_i's are the same.

To see why, we simply have to examine the alternative: if $a_1 > a_2$, then (1.2) fails: $a_1 > \max\{a_2, a_3\}$.

1.2 Four point condition: interpretation

In Figure 1, we show what the four point condition means for a distance matrix with four data points. The terms inside the similar geometric figures must add in such a way that the two largest sums agree:

$$\mathcal{D}(i,j) + \mathcal{D}(m,n) \leq \max\{\mathcal{D}(i,m) + \mathcal{D}(j,n), \mathcal{D}(j,m) + \mathcal{D}(i,n)\}.$$

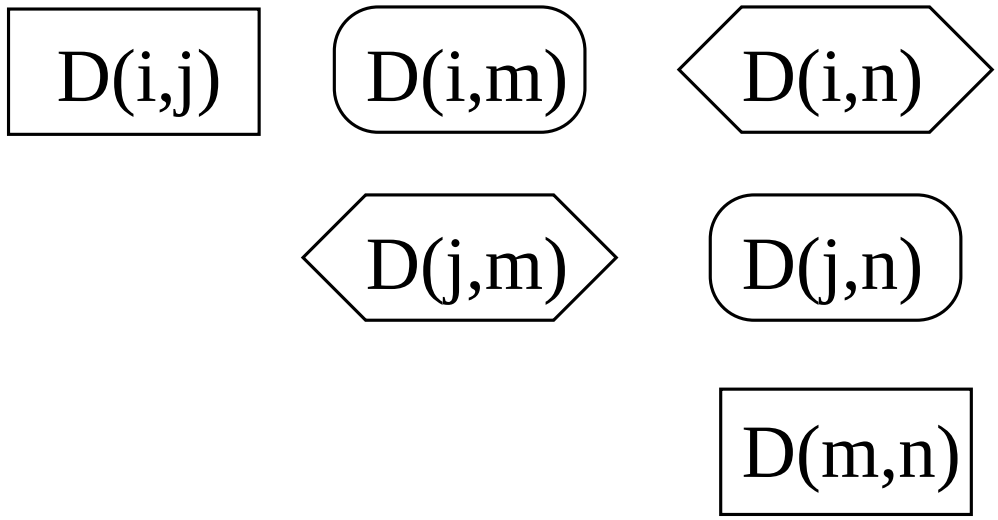

Figure 1: Entries in the distance matrix which are constrained by the four point condition.

1.3 Four point condition illustration

The distance matrix associated with the tree in Figure 2 is

$\mathcal{D}(i,j) = a + b$, $\mathcal{D}(m,n) = d + e$, $\mathcal{D}(i,m) = a + c + d$,

$\mathcal{D}(j,n) = b + c + e$, $\mathcal{D}(i,n) = a + c + e$, $\mathcal{D}(j,m) = b + c + d$.

Then

$C = \mathcal{D}(i,j) + \mathcal{D}(m,n) = a + b + d + e$,

$B = \mathcal{D}(i,m) + \mathcal{D}(j,n) = a + b + 2c + d + e$,

$A = \mathcal{D}(i,n) + \mathcal{D}(j,m) = a + b + 2c + d + e$.

Thus we see that $A = B$ and $B - C = 2c > 0$.

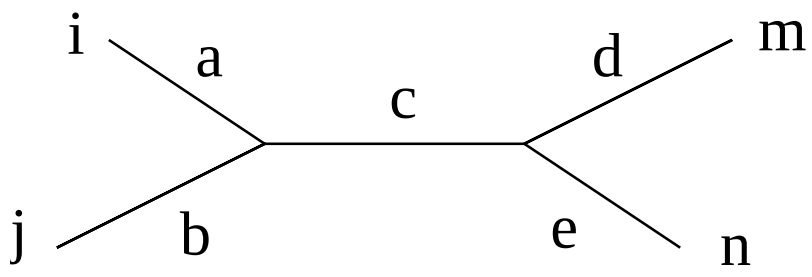

Figure 2: Configuration of tree connecting four points.

1.4 Four point condition: derivation/motivation

It is possible to motivate the derivation of the four point condition as follows. Consider the following recursive algorithm for constructing a tree from a distance matrix.

Consider any pair i, j of nodes in the discrete space.

Suppose nodes i and j are to be leaves of an internal parent node in the tree, call it k. Then define

$$\mathcal{D}(m,k) = \mathcal{D}(k,m) = \tfrac{1}{2}\left(\mathcal{D}(i,m) + \mathcal{D}(j,m) - \mathcal{D}(i,j)\right). \qquad (1.3)$$

for all the other nodes m.

Create a new discrete space by eliminating i and j and adding k; in terms of the distance matrix, we eliminate the i and j rows and add the new information defined by (1.3).

By the triangle inequality, this new matrix is non-negative.

If we find $\mathcal{D}(k,m) = 0$ we can take $k = m$ and avoid the addition to the discrete space, so we can assume that this new matrix is non-degenerate.

1.5 Is the new matrix a metric?

So far, we have not identified the distances $\mathcal{D}(i,k)$ and $\mathcal{D}(j,k)$. But of course we must have $\mathcal{D}(i,k) + \mathcal{D}(j,k) = \mathcal{D}(i,j)$. Furthermore, since every path from i and j must go through k, we must have

$$\mathcal{D}(x,y) = \mathcal{D}(x,k) + \mathcal{D}(k,y) \tag{1.4}$$

where x stands for either i or j and y stands for either m or n. Adding the four equations for the various values of x and y (and dividing by two) we get

$$\begin{aligned} \tfrac{1}{2}\big(\mathcal{D}(i,m)+\mathcal{D}(i,n) + \mathcal{D}(j,m) + \mathcal{D}(j,n)\big) & \\ = \mathcal{D}(i,k) + \mathcal{D}(j,k) + \mathcal{D}(k,m) + \mathcal{D}(k,n) & \\ = \mathcal{D}(i,j) + \mathcal{D}(k,m) + \mathcal{D}(k,n) & \end{aligned} \tag{1.5}$$

So if the triangle inequality is to hold for the new matrix, we must have

$$\tfrac{1}{2}\big(\mathcal{D}(i,m) + \mathcal{D}(i,n) + \mathcal{D}(j,m) + \mathcal{D}(j,n)\big) \geq \mathcal{D}(m,n) + \mathcal{D}(i,j) \tag{1.6}$$

for all m and n.

1.6 The new matrix is a metric

If condition (1.6) holds, we claim that this new matrix satisfies the triangle inequality, that is

$$\mathcal{D}(k,m) \leq \mathcal{D}(k,n) + \mathcal{D}(n,m) \tag{1.7}$$

for all m and n. This is equivalent to (recall the definition of $\mathcal{D}(k,m)$)

$$(\mathcal{D}(i,m) + \mathcal{D}(j,m) - \mathcal{D}(i,j)) \leq (\mathcal{D}(i,n) + \mathcal{D}(j,n) - \mathcal{D}(i,j)) + 2\mathcal{D}(n,m)$$

or, by eliminating the common term $-\mathcal{D}(i,j)$ on both sides, the same as

$$\mathcal{D}(i,m) + \mathcal{D}(j,m) \leq \mathcal{D}(i,n) + \mathcal{D}(j,n) + 2\mathcal{D}(n,m) \tag{1.8}$$

But the triangle inequality for the original matrix implies this: we just add

$$\mathcal{D}(i,m) \leq \mathcal{D}(i,n) + \mathcal{D}(n,m) \tag{1.9}$$

to

$$\mathcal{D}(j,m) \leq \mathcal{D}(j,n) + \mathcal{D}(n,m) \tag{1.10}$$

to deduce that (1.8), and therefore (1.7), holds.

1.7 Metric matrix: the rest of the story

We need to verify the rest of the requirements of the triangle inequality, e.g.,

$$\mathcal{D}(m,n) \leq \mathcal{D}(m,k) + \mathcal{D}(k,n) \tag{1.11}$$

for all m and n. Recalling the definition of $\mathcal{D}(k,m)$ and $\mathcal{D}(k,n)$, we have

$$\mathcal{D}(m,k) + \mathcal{D}(k,n) = \tfrac{1}{2}\left(\mathcal{D}(m,i) + \mathcal{D}(m,j) + \mathcal{D}(n,i) + \mathcal{D}(n,j)\right) - \mathcal{D}(i,j)$$

Therefore (1.11) is equivalent to condition (1.6).

This proves that condition (1.6) implies that the new matrix is a metric.

If the pair i and j does not satisfy condition (1.6) for all m and n, then it is not possible to identify i and j as neighbors in a tree representation of the distance matrix.

If there is no pair i and j satisfying condition (1.6) for all m and n, then it is not possible to identify any tree representation of the distance matrix with leaves as nodes.

1.8 Distance to deleted points problematic

The difficulty arises in the assignment of the distances between the new point and the deleted points. Recall that we defined the new distances by

$$\mathcal{D}(m,k) = \mathcal{D}(k,m) = \tfrac{1}{2}\left(\mathcal{D}(i,m) + \mathcal{D}(j,m) - \mathcal{D}(i,j)\right).$$

If all were well, we would have

$$\mathcal{D}(i,k) = \mathcal{D}(i,m) - \mathcal{D}(m,k) = \tfrac{1}{2}\left(\mathcal{D}(i,m) - \mathcal{D}(j,m) + \mathcal{D}(i,j)\right). \tag{1.12}$$

for any m. Since m is arbitrary, we must have

$$\mathcal{D}(i,k) = \mathcal{D}(i,n) - \mathcal{D}(n,k) = \tfrac{1}{2}\left(\mathcal{D}(i,n) - \mathcal{D}(j,n) + \mathcal{D}(i,j)\right). \tag{1.13}$$

for any other node n as well. Thus

$$\mathcal{D}(i,m) - \mathcal{D}(j,m) + \mathcal{D}(i,j) = \mathcal{D}(i,n) - \mathcal{D}(j,n) + \mathcal{D}(i,j) \tag{1.14}$$

for any m and n, which is the same as saying

$$\mathcal{D}(i,m) + \mathcal{D}(j,n) = \mathcal{D}(i,n) + \mathcal{D}(j,m) \tag{1.15}$$

which gives the four-point condition.

1.9 Interpretation of 4-point condition

We derived the necessity of the condition

$$\mathcal{D}(i,m) + \mathcal{D}(j,n) = \mathcal{D}(i,n) + \mathcal{D}(j,m)$$

based on the assumption that i and j would be child nodes of a parent node in the tree representation of the matrix.

This means that we have equality of sums of the the terms in enclosed in the identical geometric figures in Figure 3 which form part of the distance matrix.

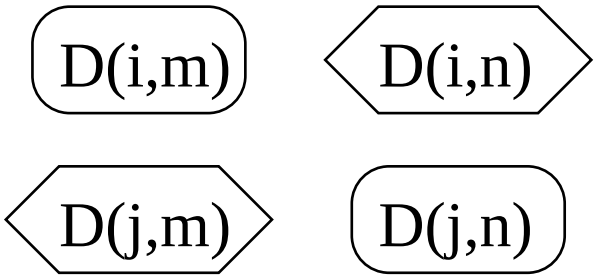

Figure 3: Entries in the distance matrix which are constrained by the four point condition.

1.10 Interpretation of 4-point condition: cont'd

The fact that

$$\mathcal{D}(i,j) + \mathcal{D}(m,n) \leq \mathcal{D}(i,m) + \mathcal{D}(j,n) = \mathcal{D}(i,n) + \mathcal{D}(j,m)$$

(as required by the four point condition) is a consequence of the fact that we assumed that i and j were nearest neighbors in the tree. This forces m and n to be nearest neighbors in the tree as well.

The common value

$$\mathcal{D}(i,m) + \mathcal{D}(j,n) - \mathcal{D}(i,j) - \mathcal{D}(m,n) = \mathcal{D}(i,n) + \mathcal{D}(j,m) - \mathcal{D}(i,j) - \mathcal{D}(m,n)$$

is twice the length of the internal edge.

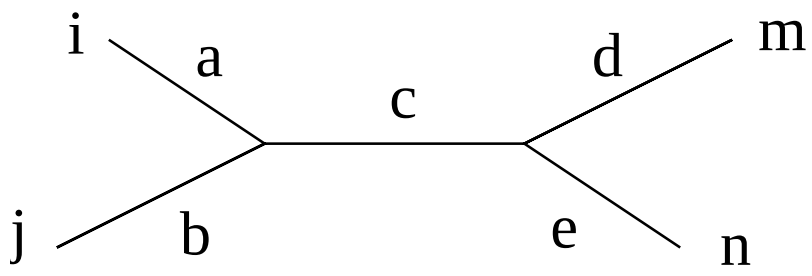

Figure 4: Configuration of tree connecting four points.

2 Example

	b:1DQJ	c:1C08	d:1NDG
a:1NDM	3	7	6
b:1DQJ		6	5
c:1C08			5

Table 1: Distance matrix for hydrogen bond interaction differences among homologous proteins: a=1NDM, b=1DQJ, c= 1C08, d= 1NDG (PDB codes).

Distance is defined as follows. First align the sequences. For each protein, define hydrogen bond matrix entry (i, j) to be one if there is an intermolecular mainchain or sidechain hydrogen bond between peptides i and j; otherwise zero.

Define the distance as the Hamming distance between the distance matrices.

For example, $D(a, b) = 3$ means that 1NDM and 1DQJ differ in exactly 3 hydrogen bonds between the antigen and antibody complex.

2.1 The tree for hydrogen bond distance

	b:1DQJ	c:1C08	d:1NDG
a:1NDM	3	7	6
b:1DQJ		6	5
c:1C08			5

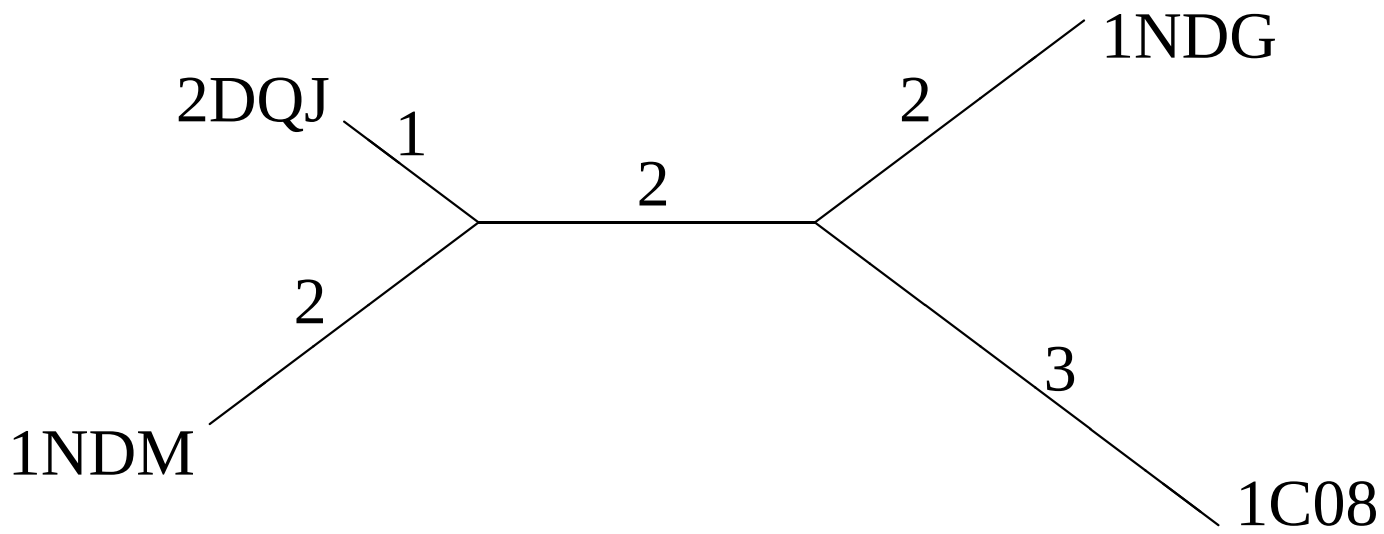

Figure 5: Tree representation of the distance matrix in Table 1.

2.2 Hydrogen bond comparisons

Following are all the intermolecular (antibody–antigen) hydrogen bonds found in the antibody complexes in the PDB files 1NDM, 1DQJ, 1C08, and 1NDG.

Can easily compute the Hamming distance

type	donor	acceptor	1NDM	1DQJ	1C08	1NDG
M-M	C-ARG 621 N	A-ASN 92 O	2.68 4.13	2.81 3.32	2.89 3.31	2.80 3.47
M-S	B-SER 354 N	C-ASP 701 OD1	3.06 7.24	3.15 6.43	3.20 6.16	(3.59 4.59)
M-S	C-GLY 702 N	B-SER 356 OG	2.88 6.46	2.82 6.33	3.00 6.21	
S-M	B-SER 331 OG	C-ARG 673 O			3.42 1.28	(4.04 0.483)
S-M	B-TYR 333 OH	C-LYS 697 O	2.73 2.97	2.64 3.49	2.69 3.85	2.63 3.60
S-M	B-TYR 350 OH	C-SER 700 O	2.55 5.35	2.60 5.02	2.64 4.63	2.66 4.46
S-M	B-TYR 358 OH	C-ASP 701 O	3.17 0.32	3.38 0.52		
S-M	C-ARG 673 NH2	B-THR 330 O		2.47 0.81		
S-S	C-ARG 673 NH1	B-THR 330 OG1	3.16 3.91			
S-S	B-SER 352 OG	C-ASP 701 OD1	2.71 8.07	2.70 7.98	2.50 10.2	2.88 2.60
S-S	B-SER 354 OG	C-ASP 701 OD1	2.85 5.56	2.49 11.5	2.77 5.93	2.80 6.64
S-S	A-GLN 53 NE2	C-ASN 693 OD1	2.85 0.56	2.84 0.85		2.81 0.82
S-S	C-ASN 693 ND2	A-GLN 53 OE1		3.30 0.228	2.83 0.375	3.30 0.296
S-S	C-LYS 696 NZ	A-ASN 31 OD1			2.95 4.76	2.83 4.64
S-S	C-LYS 696 NZ	A-ASN 32 OD1			2.82 7.13	
		total bonds	10	11	11	8

2.3 Table details

Hydrogen bond descriptors: S=sidechain, M=mainchain.

The numbers given are (1) the distance between the donor and acceptor (heavy) atoms in the hydrogen bond and (2) the quality estimate of the hydrogen bond modelled as a dipole-dipole interaction.

Note that the two bonds involving C-LYS 696 NZ in 1C08 are in conflict, in the sense that one would not normally think of the N-H group represented by NZ as capable of forming two hydrogen bonds. However, this ambiguity reflects the geometry involving this group and the two ‘acceptor’ atoms (OD1 of A-ASN 31 and A-ASN 32). In 1NDG(H8), the donor for the S-M bond with C-Arg673-O changes from B-Ser331-OG to B-Arg331-NE.

Data in parentheses are for reference only. By relaxing the definition of hydrogen bond, we can determine the interaction data for pairs of peptides not called a hydrogen bond.

3 The ABC theorem

It is possible to show that the topology of tree representations for a general distance matrix is unique under very mild conditions, as follows.

Consider the three independent quantities that figure in the four point condition:

$$\begin{aligned} A &= \mathcal{D}(i,m) + \mathcal{D}(j,n) \\ B &= \mathcal{D}(i,n) + \mathcal{D}(j,m) \\ C &= \mathcal{D}(i,j) + \mathcal{D}(n,m) \end{aligned} \tag{3.16}$$

based on the three ways to partition the index set $\{i, j, m, n\}$ into distinct pairs. These quantities determine the topology of the tree representations, as follows.

There are four distinct cases. Three of them involve two internal nodes and one internal edge, and are categorized by the following three distinct possibilities for additive matrices: $A = B > C$, $B = C > A$, and $C = A > B$. The fourth tree corresponds to $A = B = C$. Note that when $A = B = C$, the tree representing the distance matrix is a star. That is, there is one internal node k, and four edges joining the four indices to k.

3.1 Non-additive matrices

We will show that even in the case that a matrix is not additive, a unique assignment of one of these topology classes is possible in most cases.

Suppose $\mathcal{D}$ is a general distance matrix that is not necessarily additive. Without loss of generality, by renaming the indices if necessary, we can assume that the terms are ordered:

$$A \geq B \geq C. \tag{3.17}$$

The four-point condition can now be stated simply: $A = B$. In this case, the distance matrix can be represented exactly by a tree. Now we consider the other case, that $A > B$. First, we define the ℓ^1-norm for distance matrices:

$$\|\mathcal{D}\|_{\ell^1} = \sum_{i<j} |\mathcal{D}(i,j)| \tag{3.18}$$

Note that we allow for negative entries, as we intend to apply the norm to differences of distance matrices.

3.2 ABC theorem statement

The following theorem characterizes the closest additive distance matrix (one that can be represented by a tree) in the ℓ^1-norm to a general distance matrix.

Recall that, by definition, it is equivalent to say that a matrix is additive and that it satisfies the four point condition.

Theorem 3.1 *Suppose that $A > B$. Then*

$$\inf \{ \|\mathcal{D} - \mathcal{D}'\|_{\ell^1} \ : \ \mathcal{D}' \textit{ satisfies the four-point condition} \} = A - B. \tag{3.19}$$

Moreover, if $B > C$, then all additive distance matrices $\mathcal{D}'$ which satisfy

$$\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} = A - B \tag{3.20}$$

have trees with the same topology.

3.3 ABC theorem exception

When $A > B = C$, there is an ambiguity in representing $\mathcal{D}$ since there are additive matrices $\mathcal{D}'$ all equally close in ℓ^1 norm with different topology types.

We leave as an exercise to show that there is a matrix $\mathcal{D}^1$ with

$$A' = B' = C' = B = C < A,$$

as well as two others:

$\mathcal{D}^2$ with $A' = B' = A$ and $C' = B = C$

and $\mathcal{D}^3$ with $A' = C' = A$ and $B' = B = C$,

all with the property that $\|\mathcal{D} - \mathcal{D}^i\|_{\ell^1} = A - B$, but with different topologies.

(Hint: draw the different trees for the different $\mathcal{D}^i$'s.)

3.4 ABC theorem proof

To prove these assertions, we first show that

$$\inf\{\|\mathcal{D}-\mathcal{D}'\|_{\ell^1} \ : \ \mathcal{D}' \text{ satisfies four-point condition}\} \leq A-B. \tag{3.21}$$

To so so, we simply need to exhibit a $\mathcal{D}'$ which satisfies the four-point condition and $\|\mathcal{D}-\mathcal{D}'\|_{\ell^1} = A-B$. We can do this if we keep $A' = A$ and increase B' to be equal to A. For example, we can set

$$\mathcal{D}'_{in} = \mathcal{D}_{in} + A - B, \tag{3.22}$$

leaving all other entries of $\mathcal{D}'$ the same as for $\mathcal{D}$. Thus by explicit construction, we have $\|\mathcal{D}-\mathcal{D}'\|_{\ell^1} = A-B$. Similarly, since we also have $A' = A = B'$, $\mathcal{D}'$ satisfies the four point condition.

However, there is one small point that we must check. We stated that the four point condition is equivalent to $A = B$ *for a distance matrix*.

But if $\mathcal{D}'$ does not satisfy the triangle inequality, then $A = B$ is not sufficient.

3.5 ABC theorem proof, cont'd: check triangle inequality

So we need to show that $\mathcal{D}'$ satisfies the triangle inequality.

For any terms in the triangle inequality having $\mathcal{D}'_{in}$ on the right-hand side, the inequalities still hold, since we have only made the right-hand side larger. So it suffices to check that

$$\mathcal{D}'_{in} \leq \mathcal{D}'_{ix} + \mathcal{D}'_{xn} \tag{3.23}$$

for $x = j, m$. But this is equivalent to showing that (recall the definition of $A - B$)

$$\mathcal{D}_{im} + \mathcal{D}_{jn} - \mathcal{D}_{jm} = \mathcal{D}_{in} + (A - B) \leq \mathcal{D}_{ix} + \mathcal{D}_{xn} \tag{3.24}$$

for $x = j, m$. For $x = j$, this becomes

$$\mathcal{D}_{im} - \mathcal{D}_{jm} \leq \mathcal{D}_{ij} \tag{3.25}$$

which holds by the triangle inequality for $\mathcal{D}$. Similarly for $x = m$, (3.24) becomes

$$\mathcal{D}_{jn} - \mathcal{D}_{jm} \leq \mathcal{D}_{mn} \tag{3.26}$$

which also holds by the triangle inequality for $\mathcal{D}$.

3.6 ABC theorem proof, cont'd: check triangle inequality

So we need to show that $\mathcal{D}'$ satisfies the triangle inequality.

For any terms in the triangle inequality having $\mathcal{D}'_{in}$ on the right-hand side, the inequalities still hold, since we have only made the right-hand side larger. So it suffices to check that

$$\mathcal{D}'_{in} \leq \mathcal{D}'_{ix} + \mathcal{D}'_{xn}$$

for $x = j, m$. But this is equivalent to showing that (recall the definition of $A - B$)

$$\mathcal{D}_{im} + \mathcal{D}_{jn} - \mathcal{D}_{jm} = \mathcal{D}_{in} + (A - B) \leq \mathcal{D}_{ix} + \mathcal{D}_{xn}$$

for $x = j, m$. For $x = j$, this becomes

$$\mathcal{D}_{im} - \mathcal{D}_{jm} \leq \mathcal{D}_{ij}$$

which holds by the triangle inequality for $\mathcal{D}$. Similarly for $x = m$, (3.24) becomes

$$\mathcal{D}_{jn} - \mathcal{D}_{jm} \leq \mathcal{D}_{mn}$$

which also holds by the triangle inequality for $\mathcal{D}$.

3.7 ABC theorem proof, cont'd: check triangle inequality

So we need to show that $\mathcal{D}'$ satisfies the triangle inequality.

For any terms in the triangle inequality having $\mathcal{D}'_{in}$ on the right-hand side, the inequalities still hold, since we have only made the right-hand side larger. So it suffices to check that

$$\mathcal{D}'_{in} \leq \mathcal{D}'_{ix} + \mathcal{D}'_{xn}$$

for $x = j, m$. But this is equivalent to showing that (recall the definition of $A - B$)

$$\mathcal{D}_{im} + \mathcal{D}_{jn} - \mathcal{D}_{jm} = \mathcal{D}_{in} + (A - B) \leq \mathcal{D}_{ix} + \mathcal{D}_{xn}$$

for $x = j, m$. For $x = j$, this becomes

$$\mathcal{D}_{im} - \mathcal{D}_{jm} \leq \mathcal{D}_{ij}$$

which holds by the triangle inequality for $\mathcal{D}$. Similarly for $x = m$, (3.24) becomes

$$\mathcal{D}_{jn} - \mathcal{D}_{jm} \leq \mathcal{D}_{mn}$$

which also holds by the triangle inequality for $\mathcal{D}$.

3.8 ABC theorem proof, cont'd: other inequality

To prove the desired equality (3.19), must demonstrate the reverse inequality:

$$\inf \{ \|\mathcal{D} - \mathcal{D}'\|_{\ell^1} \ : \ \mathcal{D}' \text{ satisfies four-point condition} \} \geq A - B. \tag{3.27}$$

This is the same as saying that $\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} \geq A - B$ for every $\mathcal{D}'$ that satisfies four-point condition.

From the definition of the norm, we can write

$$|A - A'| + |B - B'| + |C - C'| \leq \|\mathcal{D} - \mathcal{D}'\|_{\ell^1} \tag{3.28}$$

for any distance matrices. Now suppose it were the case that for some $\mathcal{D}'$ we have $\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} < A - B$. Then we want to show that $\mathcal{D}'$ cannot satisfy the four point condition. By (3.28) we have

$$\begin{aligned} B' - A' + A - B =& A - A' + B' - B \\ \leq& \|\mathcal{D} - \mathcal{D}'\|_{\ell^1} < A - B \end{aligned} \tag{3.29}$$

from which we conclude that $B' < A'$.

3.9 ABC theorem proof: other inequality, cont'd

Similarly, since $B \geq C$, we have

$$\begin{aligned} C' - A' + A - B =& A - A' + C' - B \\ \leq& A - A' + C' - C \\ \leq& \|\mathcal{D} - \mathcal{D}'\|_{\ell^1} < A - B \end{aligned} \tag{3.30}$$

from which we conclude that $C' < A'$. So $\mathcal{D}'$ cannot satisfy the four point condition if $\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} < A - B$.

This completes the proof of the equality (3.27).

Combining (3.27) with (3.21) completes the proof of the equality (3.19):

$$\inf\{\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} \ : \ \mathcal{D}' \text{ satisfies the four-point condition}\} = A - B.$$

Now we turn to the other part of the theorem which characterizes the set of optimal distance matrices.

3.10 ABC theorem proof: optimal matrix characterization

Now suppose that $\mathcal{D}'$ is additive and satisfies (3.20):

$$\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} = A - B,$$

and $B > C$. Then we want to show that

$$A \geq A' = B' \geq B. \tag{3.31}$$

Suppose that $A < B'$. Then $B' - B > A - B$ and applying (3.28) we find that $\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} > A - B$. Therefore $A \geq B'$.

On the other hand, if $A' < B$ then $A - A' > A - B$, and again (3.28) yields a contradiction. Therefore $A' \geq B$.

If $A' < B'$, then

$$A - A' + B' - B = A - B + B' - A' > A - B \tag{3.32}$$

contradicting optimality, again via (3.28), so therefore $A' \geq B'$.

We are almost done with the proof of (3.31), but there is one more inequality to establish, namely that $A' > B'$ cannot hold.

3.11 ABC theorem proof: last step

Finally, if $A' > B'$, then the four point condition implies that $A' = C'$. Then

$$A - A' + C' - C = A - C > A - B \tag{3.33}$$

by our assumption that $B > C$, so again (3.28) yields a contradiction, implying $A' = B'$, concluding our proof that (3.31) has to hold.

Applying (3.31) in (3.28), we get

$$\begin{aligned} A - B =& A - A' + B' - B \\ =& |A - A'| + |B - B'| \\ \leq& |A - A'| + |B - B'| + |C - C'| \\ \leq& \|\mathcal{D} - \mathcal{D}'\|_{\ell^1} = A - B \end{aligned} \tag{3.34}$$

which means that equality holds throughout the experession (3.34), so we must have $C = C'$. In particular, we conclude that $A' = B' \geq B > C = C'$.

Now it is easy to show that all additive matrices with $A' = B' > C'$ have the same topology. Q.E.D.

4 What UPGMA does

UPGMA applied to a distance matrix will coallesce the closest points, that is, the one for which the entry in the distance matrix is smallests, say $\mathcal{D}(i, j)$.

One might hope that UPGMA would find the correct tree for an additive matrix.

The nearest neighbors in the tree for an additive matrix are the indices that combine to form the term C that is the smallest of the three terms involved in the four point condition: $C < B = A$.

However, even though it may hold that $\mathcal{D}(i, j)$ is the smallest entry in the distance matrix, $C = \mathcal{D}(i, j) + \mathcal{D}(m, n)$ is not smaller than A or B.

In this case, UPGMA finds the wrong tree.

For example, consider a distance matrix with $A = 5 + 5$, $B = 4 + 4$, and $C = 3 + 7$:

	b	c	d
a	3	5	4
b		4	5
c			7

Table 2: Additive distance matrix for which UPGMA gives the wrong tree.

Figure 6: UPGMA tree and additive tree for distance matrix in Table 2.

References

[1] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, January 1997.