**Course: Advanced Bioinformatics**

**Module title: Distance Based Methods**

**Module no. : 47**

**Distance Methods**

The distance method for construction of phylogenetic trees looks at the number of changes between each pair in a group of sequences to produce a phylogenetic tree of the group. The goal of distance methods is to identify a tree that positions neighbors correctly and that also has branch lengths which reproduce the original data as closely as possible.

CLUSTALW uses the neighbor-joining method as a guide to multiple sequence alignments. The PHYLIP suite of programs employ neighbor-joining methods.

Phylip <http://evolution.genetics.washington.edu/phylip.html>

Distance analysis programs in PHYLIP

FITCH: estimates a phylogenetic tree assuming additivity of branch lengths using the Fitch-Margoliash method.

KITSH: same as FITCH, but under the assumption of a molecular clock.

NEIGHBOR: estimates phylogenies using the neighbor-joining (no molecular clock assumed) or unweighted pair group method with arithmetic mean (UPGMA) (molecular clock assumed).

For phylogenetic analysis, the distance score counted as either the number of mismatched positions in the alignment or the number of sequence positions that must be changed to generate the other sequence is used.

The success of distance methods depends on the degree to which the distances among a set of sequences can be made additive on a predicted evolutionary tree.

Consider the alignment:

**A ACGCGTTGGGCGATGGCAAC**

**B ACGCGTTGGGCGACGGTAAT**

**C ACGCATTGAATGATGATAAT**

**D ACACATTGAGTGATAATAAT**

The distances between these sequences can be shown as a table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | A | B | C | D |
| A | - | 3 | 7 | 8 |
| B | - | - | 6 | 7 |
| C | - | - | - | 3 |
| D | - | - | - | - |

Using this information, an unrooted tree showing the relationship between these sequences can be drawn:

C

D

A

B

4

1

2

2

1

***Fitch and Margoliash Method***

The Fitch and Margoliash method uses a distance table. The sequences are combined in threes to define the branches of the predicted tree and to calculate the branch lengths of the tree.

Example using three sequences:

1. Draw an unrooted tree with three branches originating from a common node and label the ends:

C

c

b

a

B

A

1. Calculate the lengths of tree branches algebraically:

|  |  |  |  |
| --- | --- | --- | --- |
|  | A | B | C |
| A | -- | 22 | 39 |
| B | -- | -- | 41 |
| C | -- | -- | -- |

distance from A to B = a + b = 22 (1)

distance from A to C = a + c = 39 (2)

distance from B to C = b + c = 41 (3)

subtracting (3) from (2) yields:

b + c = 41

-a – c = -39

\_\_\_\_\_\_\_\_\_\_

b – a = 2 (4)

adding (1) and (4) yields 2b = 24; b = 12

so a + 12 = 22; a = 10

10 + c = 39; c = 29

C

29

12

10

B

A

***Example of Fitch-Margoliash Using Five Sequences***

The Fitch-Margoliash algorithm can be extended to three or more sequences. Consider the following table of distances between five separate sequences:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** | **E** |
| **A** | **--** | **22** | **39** | **39** | **41** |
| **B** | **--** | **--** | **41** | **41** | **43** |
| **C** | **--** | **--** | **--** | **18** | **20** |
| **D** | **--** | **--** | **--** | **--** | **10** |
| **E** | **--** | **--** | **--** | **--** | **--** |

Suppose that the initial tree is as follows:

# A

# B

# C

# D

# E

# a

# b

# d

# c

# e

# f

# g

1) The first step is to locate the most closely related sequences in the distance table. In this case, that would be sequences D and E.

2) Now create a new table by combining the remaining sequences. For the distance from D to A,B,C take the average distance of each of these to D ( (39 + 41 + 18) / 3 = 32.7)

For the distance from E to A,B,C, take the average distance of each of these to E ((41+43+20)/3 = 34.7). The resulting table is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **D** | **E** | **AVG ABC** |
| **D** | **--** | **10** | **32.7** |
| **E** | **--** | **--** | **34.7** |
| **AVG ABC** | **--** | **--** | **--** |

1. The average distances from D to ABC and E to ABC could also be found by averaging the sum of the appropriate branch lengths:

D to E: d + e = 10 (1)

D to ABC: d + m = 32.7 where m = g + (c + 2f + a + b) / 3 (2)

E to ABC: e + m = 34.7 (3)

By subtracting the third equation from the second equation we get:

d – e = -2

Adding this result to (1) we get: 2d = 8; d = 4

Substitute back in to get e = 6

1. Now treat D and E as a single sequence, and create a new distance table. The distance to DE is taken as the average of sequence A to D and A to E. The other distances are calculated in a similar fashion. The resulting distance table is:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **(DE)** |
| **A** | **--** | **22** | **39** | **40** |
| **B** | **--** | **--** | **41** | **42** |
| **C** | **--** | **--** | **--** | **19** |
| **(DE)** | **--** | **--** | **--** | **--** |

1. Identify the closely related sequences in the table. In this case, it is C to DE. Using algebra, the distance c can be calculated to be 9, and g is calculated to be 5.
2. Repeat the process until all lengths have been identified, in which case there is only single composite node left.

**Summary of Fitch-Margoliash Algorithm**

1. Find the mostly closely related pairs of sequences (A, B).
2. Treat the rest of the sequences as a composite. Calculate the average distance from A to all others; and from B to all others.
3. Use these values to calculate the length of the edges a and b.
4. Treat A and B as a composite. Calculate the average distances between AB and each of the other sequences. Create a new distance table.
5. Identify next pair of related sequences and begin as with step 1.
6. Subtract extended branch lengths to calculate lengths of intermediate branches.
7. Repeat the entire process with all possible pairs of sequences.
8. Calculate predicted distances between each pair of sequences for each tree to find the best tree.