As fast multiple alignment (MA) algorithms become a reality, analysis and application of their results becomes the central problem of genome research. In this book, we discuss the network structure theory of the multi-sequences induced by mutations.

6.1 General Method of Constructing the Phylogenetic Tree

6.1.1 Summary

One of the main purposes of making multiple alignments is to construct the phylogenetic tree. Looking at the MA results, we find that it is a set of sequences of the same length. If the result is correct, then this output is a kind of family file of these multiple sequences, containing all the connections among this family and the phylogenetic information on this family. Based on this family file, we may determine the evolutionary state of each sequence in this family. Generally, we use a topological tree to describe the connection among the multiple sequences, which is called a phylogenetic tree.

Tree is a class of spacial point-line graphs. The point-line graph is given by $G = \{M, V\}$, where $M = \{1, 2, \cdots, m\}$ are the points of the graph, and $V$ is the set of all pairs of points in $M$. Each pair in $V$ is seen as an arc. A point-line graph $G = \{M, V\}$ is called an undirected graph if the pairs $(s, t), (t, s) \in M$ are the same. Otherwise, it is a directed graph. These two types of point-line graphs will frequently appear in the following text. The point-line graphs theory is considered in many books, and it is not discussed further in this book.

There are many methods for constructing the phylogenetic tree. We will introduce these methods in this section as follows:

1. Distance-based methods (e.g., neighbor-joining). Any alignment result may be used to compute a distance matrix between these sequences. Based
on this distance matrix, we may produce the corresponding phylogenetic tree. The most popular methods are called UPGMA and neighbor-joining.

2. Feature-based methods (e.g., maximum parsimony method). This kind of method uses the features (characteristics) of the alignment outputs to construct the phylogenetic tree.

3. Probability-based methods (e.g., maximum-likelihood method and Bayes method). Using these methods to construct the phylogenetic tree, we should begin by constructing a probability model for the sequence mutation, and then construct the phylogenetic tree based on both the output and the probability model.

6.1.2 Distance-Based Methods

There are many distance-based methods for constructing the phylogenetic tree, and we only introduce two of these in this subsection, namely, UPGMA and neighbor-joining.

Unweighted Pair Group Method with Arithmetic Mean

Unweighted pair group method with arithmetic mean (UPGMA) [63,96] is the simplest of all clustering methods used to construct a phylogenetic tree. This method requires that the substitution velocity of the nucleotides or amino acids be uniform and unchanging through the entire evolution process. In other words, the molecular clock hypothesis holds. At each parent node, the branch lengths from the parent node to the two child nodes are the same.

The most intuitive clustering method used to construct the phylogenetic tree is the system clustering method. This method assembles the two nearest classes to a new class, into a cluster each time, until all the classes are assembled into one class. The algorithm is trivially developed by following the steps listed below:

1. Given an $n$-multiple nucleotide sequence or amino acid sequence, choose a distance function (e.g., using the Hamming distance function) and compute the evolution distance for every pair of sequences based on their pairwise alignment result, producing a distance matrix.
2. Regard each sequence as a class, then use the $n$ initial classes as the leaf-nodes of the phylogenetic tree.
3. Using the distance matrix, search the two classes $X,Y$ that are nearest, and then assemble $X,Y$ into a new class $Z$, which is then the parent node of $X,Y$. The distances from node $Z$ to $X$ and to $Y$ (that is, the branch lengths from $Z$ to $X$ and to $Y$) are the same, and equal to $d(X,Y)/2$. The total number of classes is then $n - 1$.
4. Compute the distances from the new node $Z$ to other nodes. Let $K$ be the query node for the distance to be computed from $K$ to $Z$. Since $d(X,K)$
and $d(Y, K)$ are collected in the distance matrix, we compute the distance $d(Z, K)$ by one of the following ways:

\[
\begin{align*}
    d(Z, K) &= \min\{d(X, K), d(Y, K)\}, \\
    d(Z, K) &= \max\{d(X, K), d(Y, K)\}, \\
    d(Z, K) &= (d(X, K) + d(Y, K))/2 .
\end{align*}
\]

We then find a new distance matrix.

5. Repeat steps 3 and 4 until all the classes are assembled into one.

This clustering method is easy to operate. In fact, this procedure is simply a MA process, and the result involves making MA using the pairwise alignment algorithm, based on the multiple sequences.

UPGMA is used to construct the phylogenetic tree in a way similar to the system clustering method, the main difference being the formula used to compute the distance of classes. Using step 4 above to compute the distance between two classes, if the numbers of the sequences in the two classes are different, we have to compute the distance from the new cluster to all other clusters as a weighted average of the distances from its components:

\[
d(Z, K) = \frac{n_X}{n_X + n_Y} D(X, K) + \frac{n_Y}{n_X + n_Y} D(Y, K) ,
\]

where $n_X$ and $n_Y$ are the number of sequences in $X$ and $Y$, respectively.

The Neighbor-Joining Method

The neighbor-joining method [81] is a distance-based method used to construct a phylogenetic tree. This method does not depend on the molecular clock hypothesis, and it can process large-size sequences quickly. It has therefore been a popular method for constructing phylogenetic trees up to now.

Neighbor-joining is also a clustering method. We can prove that the summation of all the branch lengths in the phylogenetic tree generated by this method is the smallest. The phylogenetic tree with the smallest sum of branch lengths is not unique, but this method produces only one.

The neighbor-joining method starts from a starlike structure, and collects all “neighbors” together to form a tree without roots as the output. For a set of $N$ sequences, the computing steps are given as follows:

1. Compute the distance matrix of the $N$ sequences with respect to some chosen metric.
2. Regarding each sequence as a node, the initial topological structure is starlike, as in the schematic representation shown in Fig. 6.1a.
3. For an arbitrary pair of nodes, we compute the sum of all branch lengths if we combine this pair of nodes as a new node. Let $D_{ij}$ be the distance between sequences $i$ and $j$, and this distance can be obtained from step 1;
$L_{ab}$ is the length between node $a$ and node $b$, then the sum of the branch lengths of the starlike structure (Fig. 6.1a) is defined as follows:

$$S_0 = \sum_{i=1}^{N} L_{iX} = \frac{1}{N-1} \sum_{i<j} D_{ij} , \quad (6.1)$$

where $X$ is the only inner node at the center of the starlike structure. The $\frac{1}{N-1}$ in formula (6.1) is due to the fact that each edge is counted $N - 1$ times. We may assume that nodes 1 and 2 are joined. As in Fig. 6.1b, nodes 1 and 2 are seen as one class, and the other nodes as another class. The inner nodes are $X$ and $Y$ and the branch length $L_{XY}$ between $X$ and $Y$ is defined by

$$L_{XY} = \frac{1}{2(N-2)} \left( \sum_{k=3}^{N} (D_{1k} + D_{2k}) - (N-2)(L_{1X} + L_{2X}) - 2 \sum_{k=3}^{N} L_{iY} \right) , \quad (6.2)$$

where the first term in parentheses is the sum of the lengths from the other nodes to nodes 1 and 2. The latter two terms are irrelevant to $L_{XY}$ and should be subtracted because $L_{XY}$ is counted $2(N-2)$ times in the first term in parentheses. Following Fig. 6.1b and the definition of the branch length, we have

$$L_{1X} + L_{2X} = D_{12} , \quad \sum_{k=3}^{N} L_{iY} = \frac{1}{N-3} \sum_{3\leq i<j} D_{ij} , \quad (6.3)$$

and

$$S_{12} = L_{XY} + (L_{1X} + L_{2X}) + \sum_{k=3}^{N} L_{iY} . \quad (6.4)$$

Making use of (6.2) and (6.3), we have

$$S_{12} = \frac{1}{2(N-2)} \sum_{k=3}^{N} (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N-2} \sum_{3\leq i<j} D_{ij} , \quad (6.5)$$
in which \( D_{ij} \) are known. Therefore, following from (6.5), we may compute the sum of the branch lengths if nodes 1 and 2 are joined. Similarly, if an arbitrary pair of nodes are joined, we can compute the corresponding sum of the branch lengths.

4. Compare all sums of the branch lengths obtained in step 3, and choose this pair of nodes as the “neighbor” in case it minimizes the sum of branch lengths. We find the topological structure shown in Fig. 6.1b if nodes 1 and 2 are joined. The branch lengths \( L_{1X} \) and \( L_{2X} \) are then computed as follows:

\[
L_{1X} = \frac{(D_{12} + D_{1Z} - D_{2Z})}{2}, \quad L_{2X} = \frac{(D_{12} + D_{2Z} - D_{1Z})}{2},
\]

(6.6)
in which \( D_{1Z} = \sum_{i=3}^{N} D_{1i}/(N - 2) \) and \( D_{2Z} = \sum_{i=3}^{N} D_{2i}/(N - 2) \).

5. Compute the distance between the new node and other nodes. We may again assume that the new node is joined by nodes 1 and 2, and the distance between the new node and the \( j \)th old node is defined as

\[
D_{(1-2)j} = \frac{(D_{1j} + D_{2j})}{2}, \quad j = 3, 4, \cdots, 8.
\]

(6.7)

Therefore, the total number of outer nodes decreases from \( N \) to \( N - 1 \), and inner nodes increase from 1 to 2.

6. Repeat steps 3–5 until the inner nodes become \( N - 3 \). We then have a tree without a root, as required. To help the reader understand this method more easily, we give an example to illustrate how to use the neighbor-joining method to construct a phylogenetic tree.

**Example 20.** Let the distance matrix of the five species \( A, B, C, D, \) and \( E \) be

\[
\begin{pmatrix}
A & B & C & D \\
B & 7 & \text{} & \text{} \\
C & 8 & 5 & \text{} \\
D & 11 & 8 & 5 \\
E & 13 & 10 & 7 & 8
\end{pmatrix}
\]

We construct its phylogenetic tree using the neighbor-joining method.

Let us compute the sum of all branch lengths when two nodes are joined using formula (6.4). Then

\[
\begin{pmatrix}
(S) & A & B & C & D \\
B & 19.33 & \text{} & \text{} & \text{} \\
C & 20.67 & 20.67 & \text{} & \text{} \\
D & 21.00 & 21.00 & 20.33 & \text{} \\
E & 21.00 & 21.00 & 20.33 & 19.67
\end{pmatrix}
\]

From this matrix, we find that \( S_{AB} = 19.33 \) is a minimum. Thus, \( A \) and \( B \) are “neighbors” and we join \( A \) and \( B \) as a class, and then add an inner node \( X \). The topological structure of the tree is shown in Fig. 6.2b. Using formulas (6.5) and (6.6), we find that \( L_{AX} \) and \( L_{BX} \) are 5 and 2, respectively. There are
then two inner nodes, so we continue the procedure. Following from formula (6.7) we find a new distance matrix as follows:

\[
\begin{pmatrix}
A - B & C & D \\
C & 6.5 \\
D & 9.5 & 5 \\
E & 11.5 & 7 & 8
\end{pmatrix}.
\]

Repeating the above process, we obtain a new matrix of the sums of branch lengths:

\[
\begin{pmatrix}
(S) & A - B & C & D \\
C & 15.5 \\
D & 16 & 16 \\
E & 16 & 16 & 15.5
\end{pmatrix}.
\]

From the above matrix, we find that the sum of the branch lengths when \(A - B\) and \(C\) are “neighbors” is the same as when \(D\) and \(E\) are “neighbors”. If \(A - B\) is seen as a node, then the topological structures of the trees for both cases are the same. Thus, let \(A - B\) and \(C\) be “neighbors”, and add a new inner node \(Z\). The tree then has three inner nodes, and the minimum distance tree appears. Following from formulas (6.5) and (6.6), we get \(L_{(A-B)Y} = 5.5\), \(L_{CY} = 1\), \(L_{DZ} = 3\) and \(L_{EZ} = 5\). Furthermore, the lengths of the other branches are computed as:

\[
L_{XY} = L_{(A-B)Y} - (L_{AX} + L_{BX})/2 = 5.5 - 3.5 = 2,
L_{YZ} = L_{CD} - L_{CY} - L_{DZ} = 1.
\]

This ends the procedure to construct the phylogenetic tree; the process is shown in Fig. 6.2.

### 6.1.3 Feature-Based (Maximum Parsimony) Methods

Feature-based methods often use the discrete features of data, for example, using alignment outputs for DNA or protein sequences to construct the phylogenetic tree. The most popular method is the maximum parsimony method, which uses features of DNA sequences to construct the phylogenetic tree.
These features of DNA sequences include the positions where the nucleotides differ. For positions where the nucleotides are the same for all sequences, the position does not join to construct the required phylogenetic tree if we use feature-based methods. However, they do join to construct the tree if we use distance-based methods. This is a major difference between feature-based and distance-based methods.

**The Outline of the Maximum Parsimony Method**

1. Perform the MA of the given multiple sequences, and obtain an output in which every sequence has the same length.
2. Based on the alignment output, we look for the informative positions. A position is defined as the informative position if at least two kinds of nucleotides occur with a high frequency in the column corresponding to this position. Otherwise, this position is a noninformative position. In the following example, the fifth, seventh, and ninth positions are informative positions marked with an asterisk, and the other positions are noninformative.

   |    | 1  | 2  | 3  | 4  | 5* | 6  | 7* | 8  | 9* |
   |----|----|----|----|----|----|----|----|----|----|
   | 1  | A  | A  | G  | A  | G  | T  | G  | C  | A  |
   | 2  | A  | G  | C  | C  | G  | T  | G  | C  | G  |
   | 3  | A  | G  | A  | T  | A  | T  | C  | C  | A  |
   | 4  | A  | G  | A  | G  | A  | T  | C  | C  | G  |

3. Construct the maximum parsimony phylogenetic tree based on the informative positions. We begin by giving all topological structures of possible phylogenetic trees for the sequences. For each of these trees, we let the informative positions be the leaf nodes, and we then predict their parent nodes based on the information of the leaf nodes, as well as giving the statistics of the differences between nucleotides within the neighbor nodes and computing the sum of the difference of nucleotides on the whole tree, which is called the length of the tree. We choose the tree with the minimum length as the estimation of the phylogenetic tree. For the above example, these four sequences may result in three possible trees without

**Fig. 6.3a–c.** Using the maximal parsimony method to construct phylogenetic tree. 

- **a** The topological structure of the first tree, whose length is 4.
- **b** The topological structure of the second tree, whose length is 5.
- **c** The topological structure of the third tree, whose length is 6.
a root as shown as Fig. 6.3. For every possible tree, we compute the number of substitutions at the informative positions. We find the lengths of the three trees to be 4, 5, and 6, respectively. Therefore, we choose the tree in Fig. 6.3a as the estimation of the phylogenetic tree.

**Calculation Using the Fitch Algorithm**

In the above case, the parent nodes are easy to identify, as is the length of the tree. For the complex case where the tree has roots, then the length of the tree is calculated using the Fitch [30] algorithm as follows:

1. Give the range of each node. We define the range of the successor node as all the nucleotides occurring in the column corresponding to the successor node. For the inner nodes, the range is defined as the intersection of the ranges of the two successor nodes if it is not empty, or the union of the ranges of two successor nodes if their intersection is empty. Therefore, we may get the ranges of all the inner nodes and successor nodes.

2. Determine the value of each node. This process is opposite to the one above. We start from the value of the parent nodes to get the value of the successor nodes. For the root node, we choose an arbitrary value from its range as the value of the root node. For an inner node, if its range includes the value of its parent node, then this common value is defined as the value of this inner node. Otherwise, we select a value randomly from the range of this inner node as its value.

3. Determine the substitution times of the tree. The substitution times for the tree are defined as the total number of times the intersection set of the ranges of all the successor nodes generated in the first step is not empty. Therefore, for a given tree with roots, we obtain the substitution times at each informative position according to the above three steps. The sum of the substitution times is the length of the tree.

We have outlined the process of constructing a phylogenetic tree using the maximum parsimony method. However, there remain some questions to be answered. First, if the number of species is too large, then the topological structures of the phylogenetic tree will generally be too high in number. For example, in trees with roots, when the number of species is $n \geq 2$, the number of trees with roots is $N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$. Therefore, the number increases exponentially. Typically, the number of trees with roots is about $3.4 \times 10^7$ if $n = 10$; and the number of trees with roots is about $8.2 \times 10^{21}$ if $n = 20$. This number is too large to compute the minimum length of a tree. Therefore, we must attempt to decrease the search times. For example, the branch and bound algorithm ensures a minimum length tree will be found. However, the time complexity of the algorithm is close to that of an exhaustive search algorithm in the worst case scenario. It is a time-consuming method. Heuristic search algorithms are another option. They highly reduce the search times but do not ensure the optimal solution will be found. Therefore, we consider
using exhaustive search algorithms or the branch and bound algorithm only as long as the number of species is not excessive. It may be worth using heuristic search algorithms if the number of species is high. In addition, repeating this algorithm as the order of species changes will be helpful towards improving the quality of the result.

The second question is in regard to the probability of different nucleotide substitution in a true evolution process. For example, the number of transversion mutations is larger than the number of transition mutations in the real evolutionary process. This reminds us that the transition and transversion mutations are not equal. This question was not addressed in Fitch’s algorithm. However, Sankoff’s algorithm [82] offers a solution to this problem. The algorithm can deal with multiple features, and discusses the difference in probability corresponding to different features.

The maximum parsimony method uses the information on all the nucleotides at the informative positions to construct the phylogenetic tree. The advantages of this method are as follows: It uses the information on the alignment output completely. It obtains the information of ancestor sequences, and it does not show the difference between the nucleotides as is the case with the distance-based method. However, its disadvantages are also significant in that it does not use the information on the noninformative positions, its speed is much longer than that of distance-based methods, and the phylogenetic tree does not offer information about branch lengths. These weaknesses limit its applications.

6.1.4 Maximum-Likelihood Method and the Bayes Method

Among all the methods for constructing the phylogenetic tree, the maximum-likelihood method and the Bayes method are currently the most popular [2,28, 44,108,110,112]. These two methods are based on the use of probability theory to estimate the most probable topological structure of the phylogenetic tree. It allows different positions of sequences and different periods with evolution rate. It is the most credible method for constructing the phylogenetic tree. The well-known system analysis software programs PAML and MrBayes utilize the maximum-likelihood-based method and the Bayes inference-based method, respectively.

The Probability Models for Evolution

In Chap. 1, we introduced types-I, type-II, type-III, and type-IV mutations. For the conservation sequences, the probability that type-II, type-III, and type-IV happen in these sequences is small enough that we may ignore it. In other words, we will not consider this position if there is an insertion or deletion happening at this position. We consider type-II mutations to have the same effect as a type-I mutation occurring twice. For simplicity, we assume that only type-I mutations are occurring in these sequences.
We focus the discussion on DNA sequences and let the DNA sequence be of the following form:

\[ A_t = (a_{t_1}, a_{t_2}, \ldots, a_{t_m}), \quad t \in \mathbb{R}, a_{t_j} \in \mathbb{Z}_4, \]

(6.8)

where \(m\) is the length of the sequence. Let \(A_0\) be the ancestor sequence and let \(A_t\) be the state that the ancestor sequence evolves to at the time \(t\). The state at the \(j\)th position of \(A_t\) is considered a random variable \(\xi_{t_j}\). For a given position \(j\), the sequence \(\{a_{t_j}, t \in [0, +\infty]\}\) is seen as a trail of the stochastic process \(\{\xi_{t_j}, t \in [0, +\infty]\}\). We assume that the evolutions of the sequence are independent; in other words, that \(\xi_{t_j}\) is independent of \(j\). That is to say that we only consider evolution at the \(j\)th position. For simplicity, we write \(\{\xi_{t_j}, t \in [0, +\infty]\}\) as \(\{\xi_t, t \in [0, +\infty]\}\).

We assume that the evolution process is a homogeneous Markov process, i.e., for any \(t \geq 0\), the conditional probability

\[ p_{YX}(t) = P\{\xi_{t+s} = Y | \xi_s = X\} \]

(6.9)

does not depend on \(s (s \geq 0)\) where \(p_{YX}(t)\) is the transition probability of \(\xi\) from state \(X\) to state \(Y\) after time \(t\). Note that events at time \(t\) and at time \(s\) are independent, and following from the C-K equation, we obtain

\[ p_{YX}(t+s) = \sum_{z \in \mathbb{Z}_4} p_{YZ}(t) p_{ZX}(s). \]

(6.10)

If we know the ancestor sequence of the Markov process, i.e., if \(\xi_0\) is given, the process is unique if we get the transition probability matrix of the Markov process \(P(t) = (p_{YX}(t))_{4 \times 4}\) where this transition probability matrix is the so-called substitution matrix. For example, to analyze the evolution of a protein, the PAM matrix and BLOSUM matrix are well-known, and these are the transition probability matrices we will discuss. The identifier numbers 0, 60, and 250 following the letters PAM in the matrices PAM0, PAM60, and PAM250 simply correspond to the \(t\) in the transition probability matrix \(P(t)\), which is the evolution time.

To obtain \(P(t)\), we assume that the following relationship holds:

\[ \lim_{t \to 0^+} p_{YX}(t) = \delta(Y, X) = \begin{cases} 1, & Y = X, \\ 0, & Y \neq X. \end{cases} \]

(6.11)

This assumption indicates the probability that \(\xi\) was substituted in a very short time is 0, i.e., \(P(0) = I\), where \(I\) is a \(4 \times 4\) unit matrix. Let \(Q\) be the right derivative matrix of \(P(t)\) at \(t = 0\), then

\[ Q = P'(0) = \lim_{t \to 0^+} \frac{P(t) - I}{t}, \]

(6.12)

namely,

\[ P(dt) = Q dt + I. \]

(6.13)
From (6.13) we get

\[ P(t + dt) = P(t)P(dt) \]  

(6.14)

From (6.14), we replace \( P(dt) \) with \( Qdt + I \) on the right side, to get

\[ P(t + dt) - P(t) = P(t)Qdt \]

namely,

\[ P'(t) = QP(t) \]  

(6.15)

Solving the differential equation, we find

\[ P(t) = e^{tQ} = I + \sum_{n=1}^{\infty} \frac{Q^n t^n}{n!} \]  

(6.16)

This is the transition probability matrix we require. Using this formula, we find that the transition probability matrix is uniquely determined by the right derivative matrix \( Q \) of \( P(t) \) at \( t = 0 \) where \( Q \) is the so-called instantaneous transition probability matrix. If \( Q \) is symmetrical, then \( P \) is also symmetrical. This means the evolution process is reversible. If \( Q \) is an arbitrary matrix, then the formula (6.16) can be difficult to compute.

In practice, homogeneity, stationarity, and reversibility of Markov processes are all required. Homogeneity in the evolution process is equivalent to \( Q \) being independent over time. Stationarity in the evolution process means that the percentage of the nucleotides in the sequence is unchanged. Reversibility is obeyed when \( \pi_X \Pi_{XY}(t) = \pi_Y \Pi_{YX}(t) \) holds, where \( \pi_X \) is the percentage of the nucleotide \( X \) in the sequence. This means that in theory we cannot distinguish a forward process from a reverse process. In a reversible process, we can diagonalize the matrix \( Q \), i.e., it can be decomposed as \( U \cdot \text{diag}\{\lambda_1 t, \ldots, \lambda_4 t\} \cdot U^{-1} \), where \( \{\lambda_1, \ldots, \lambda_4\} \) is the characteristic vector of \( Q \). Thus, the formula (6.16) may be readily computed as follows:

\[ P(t) = e^{tQ} = I + \sum_{n=1}^{\infty} \frac{Q^n t^n}{n!} = U \cdot \text{diag}\{e^{\lambda_1 t}, \ldots, e^{\lambda_4 t}\} \cdot U^{-1} \]  

(6.17)

The whole evolution process is determined with the computation of \( P(t) \). This probabilistic model is supported by the three following suppositions:

1. This evolution process only involves type-I mutations.
2. The evolution processes at every pair of positions are independent.
3. The evolution process is an homogeneous, stationary, and reversible Markov process at each position.

In practice, the evolution process is not so ideal; insertions and deletions may happen although the sequences are conserved. These assumptions have little effect on the result.

The above evolution model is idealized, which tells us that the evolution process is determined by its initial state. In other words, the evolution process
is determined by $Q$ which is the right derivative of the transition probability matrix at time 0, or the instantaneous transition probability matrix. The matrix $Q$ depends on the ancestor sequence. In practice, however, we know the present sequences, not the ancestor sequences. If we have the instantaneous transition probability matrix $Q$ and the present sequence, we may predict the ancestor sequence and construct the entire phylogenetic tree.

Maximum-Likelihood Method for Constructing the Phylogenetic Tree

On one hand, the whole evolution process is determined by the instantaneous transition probability matrix $Q$ according to the probabilistic evolution model. On the other hand, the probabilities of the phylogenetic tree may be computed if the topological structure of a phylogenetic tree is given. Therefore, for multiple sequences, we may use a maximum-likelihood method to get a maximum probability phylogenetic tree. This can be considered the maximum likelihood estimate of the true phylogenetic tree.

We assume that the probability of substitutions happening over an infinitesimal time interval $\Delta t$ is $\lambda \Delta t$. Let the probability that the nucleotide mutates to $X$ be $p_X$. Then, within $\Delta t$, the probability that $X$ mutates to $Y$ is

$$p_{XY}(\Delta t) = \begin{cases} 1 - \lambda \Delta t, & \text{if } X = Y, \\ \lambda \Delta t p_Y, & \text{otherwise} \end{cases}.$$  

(6.18)

Following from the definition of $\delta(Y,X)$ given in the last section, we have

$$p_{XY}(\Delta t) = (1 - \lambda \Delta t)\delta(Y,X) + \lambda \Delta t p_Y.$$  

(6.19)

Since the number of substitutions obeys the Poisson distribution, for a small $t$, $e^{-\lambda t}$ is the probability that there is no substitution happening within $(0, t)$. Thus, the above formula can be corrected as follows:

$$p_{XY} = e^{-\lambda t} \delta(Y,X) + (1 - e^{-\lambda t})p_Y.$$  

(6.20)

Generally, the distribution $p$ is the stationary distribution of the Markov process if it is stationary. Based on the alignment output for multiple sequences, we may use the percentage of each nucleotide as the estimation of the stationary distribution. We may then evaluate the probability that nucleotide $X$ mutates to $Y$ within an interval $(0, t)$.

In conclusion, if multiple sequences are given, we can obtain the alignment output. At each position, we choose a proper parameter $\lambda$, and choose a topological structure of the tree and the sum of branch lengths, and then we may find the probability to generate the phylogenetic tree at this position. This routine is shown in Fig. 6.4.

There are four species on the phylogenetic tree without roots. The length of the branches is measured by the average numbers of nucleotides substituted at this position $\{v_i, i = 1, 2, 3, 4, 5\}$. 
6.1 General Method of Constructing the Phylogenetic Tree

Fig. 6.4. The topological structure of a phylogenetic tree for four species. (From [108])

We may assume that the length of the alignment output for the four species is \( n \) where we ignore the insertion and deletion, i.e., neither type-III nor type-IV mutations happen. If there is an insertion or deletion at one position, the column corresponding to this position will be deleted. Let the nucleotides at the \( h \)th position of the MA output be \( x_h = \{x_1, x_2, x_3, x_4\}^T \) and let \( \{\pi_i, i = 1, 2, 3, 4\} \) be the stationary distribution of nucleotides, which can be approximated by the percentage of each nucleotide. Therefore, to generate the phylogenetic tree as in Fig. 6.4, the probability at position \( h \) is computed in the following way:

\[
P(x_h, v) = \sum_{x_5=1}^{4} \sum_{x_6=1}^{4} \pi_{x_5} (P_{x_5x_1}(v_1)P_{x_5x_2}(v_2)P_{x_5x_6}(v_5) \times P_{x_6x_3}(v_3)P_{x_6x_4}(v_4)) .
\]

(6.21)

If the molecular clock supposition holds, then the formula for the probability to construct the phylogenetic tree holds for any position. However, in most cases, the molecular clock supposition does not hold. The evolution speeds are different as the position is changed. That is, at different positions, the same branch length may not represent the same evolution time or the same substitution numbers. Therefore, \( \lambda \) is connected with the positions. As a result, Yang [108] proved that the distribution of \( \lambda \) is approximated by a \( \Gamma \) distribution. Let the value of \( \lambda \) at position \( h \) be \( \lambda_h \) so that the above formula can be written as

\[
P(x_h, v|\lambda_h) = \sum_{x_5=1}^{4} \sum_{x_6=1}^{4} (\pi_{x_5}(P_{x_5x_1}(v_1\lambda_h)P_{x_5x_2}(v_2\lambda_h)P_{x_5x_6}(v_5\lambda_h)) \times P_{x_6x_3}(v_3\lambda_h)P_{x_6x_4}(v_4\lambda_h)) ,
\]

(6.22)

where \( P \) can be obtained from formula (6.22).

Furthermore, we assume that evolutions at different positions are independent. The probability that the whole sequence generates Fig. 6.4 is then computed by following formula:

\[
P(X|T) = \Pi_{h=1}^{n} E(P(x_h, v|\lambda_h)) ,
\]

(6.23)
where the expectation value at the right side of the equation is under the condition $\lambda_h$, $T$ is the phylogenetic tree including branch length information. This equation is called the likelihood equation. Taking the logarithm of both sides of the equation, we get the following logarithm likelihood equation:

$$l = \sum_{h=1}^{n} \log(E(P(x_h, v|\lambda_h))).$$

(6.24)

In the above equation, we find the maximum value of $T$, and the maximum likelihood estimate of the phylogenetic tree.

Generally, nucleotide substitution involves not only stationary distribution, but also the percentages of transverse/transition mutations, and synonymous/nonsynonymous mutations. Currently, instantaneous transition probability matrices are commonly used. For example, the Jukes–Cantor model [49], F81 model, K2P model [52], HKY model, GTR model [100,109], etc. all involve this matrix.

The maximum-likelihood method to construct a phylogenetic tree gives a probabilistic view of evolution. This model is superior to others. Especially in simulation research, this method is better than feature-based methods and distance-based methods. In different regions, we can choose different instantaneous transition probability matrixes. For example, in the region that code a protein, we may use the substitution model of a codon to construct the phylogenetic tree [34], while maximum likelihood methods would be time-consuming. For large size data, this method takes too long, or may not work at all.

The Bayes Method of Constructing the Phylogenetic Tree

The Bayes method of constructing the phylogenetic tree is based on the posterior probability distribution. We use the phylogenetic tree with the maximum posterior probability as the estimation of the true phylogenetic tree. Of course, we can use the Bayes formula to compute the $P(X|T)$ that is used in the maximum likelihood method as follows:

$$P(T_i|X) = \frac{P(X|T_i)P(T_i)}{P(X)} = \frac{P(X|T_i)P(T_i)}{\sum_{T_i} P(X|T_i)P(T_i)},$$

(6.25)

where $T_i$ is the topological structure and the branch lengths of some tree, $X$ is a multiple sequence, and $P(X|T_i)$ is the conditional probability computed by formula (6.23).

Obviously, the posterior probability shown as (6.25) cannot be obtained through analytical approaches. The Monte Carlo method is a better tool to solve this problem. A popular method is the the Metropolis-Hastings method [37, 39, 62]; this is a Monte Carlo Markov chain (MCMC) method. It is outlined as follows:
1. Let $T$ be the current state of the Markov chain. For the initial state, the selection of $T$ is random.

2. Select a new state $T'$ based on the transition probability matrix of the Markov chain. Generally, this state transition probability matrix is symmetrical. The probability from state $T$ to $T'$ is equal to that from state $T'$ to $T$.

3. The probability that the new state is acceptable is computed as follows:

$$R = \min\left(1, \frac{P(X|T')}{P(X|T)} \times \frac{P(T')}{P(T)} \times \frac{q(T,T')}{q(T',T)} \right),$$

where $q$ is the transition probability matrix of the Markov chain, and $\frac{q(T,T')}{q(T',T)} = 1$ if $q$ is symmetric.

4. Generate a random number $U$ in the open interval $(0,1)$. Then let $T = T'$ if $U \leq R$, and keep the state $T$ unchanged if $U > R$.

5. Repeat steps 2–4.

The distribution of $T$ obtained from the above steps is the distribution of $T$ in (6.25). We choose the maximum probability tree as the Bayes estimation of the real phylogenetic tree. Additionally, (6.26) is easy to compute because the large denominator is canceled. Therefore, in order to construct the phylogenetic tree, we choose this method when processing large-sized sequences.

6.2 Network Structure Generated by MA

The network structure generated by the MA outputs was proposed as a generalization of graphs and trees. We show that general theory of graphs and trees is perfectly suited to the analysis of the network structure generated by MA.

6.2.1 Graph and Tree Generated by MA

As above, let $A = \{A_1, A_2, \cdots, A_m\}$ be a multiple sequence, and let $C = \{C_1, C_2, \cdots, C_m\}$ be the alignment output. We then analyze the network structure generated by $C$.

The Data Structure Generated by MA

The various data structures generated by the MA output $C$ are defined as follows:

1. The distance matrix generated by MA is defined as:

Let $D = (d_{s,t})_{s,t=1,2,\cdots,m}$, where $d_{s,t} = d(C_s, C_t) = \sum_{j=1}^{n'} d(c_{s,j}, c_{t,j})$ and $d(c,c')$, $c,c' \in V_{q+1}$ be the distance function defined on $V_{q+1}$. Then $\hat{M} = \{M, D\}$ is a metric space.
Remark 5. The definition involves the alignment output \( C \), while it is not necessary for \( C \) to be the optimal alignment of \( A \).

2. Stable and unstable regions: A given \( j \) is the stable position if \( c_{1,j} = c_{2,j} = \cdots = c_{m,j} \) holds. Otherwise, this position is an unstable position. A region is stable if all positions in this region are stable, and a region is unstable if all the positions in this region are not stable. Let \( \Delta_0 \) and \( \Delta_1 \) be the stable region and unstable region of \( C \), respectively.

The definition of a stable region and an unstable region can be generalized to the partial alignment case. Let \( M_0 \) be a subset of \( M \), then \( C_0 = \{c_s, s \in M_0\} \) is the partial sequence of \( C \). With this new set, we may divide \( N' = \{1, 2, \cdots, n'\} \), the set of positions of \( C \), into three parts as follows:

\[
\begin{align*}
\Delta_0(M_0) &= \{j \in N, c_{s,j} = c_{s',j} \neq q, \forall s, s' \in M_0\}, \\
\Delta_1(M_0) &= \{j \in N, \text{there is a pair } s \neq s' \in M_0, \text{such that } c_{s,j} \neq c_{s',j}\}, \\
\Delta_2(M_0) &= \{j \in N, c_{s,j} = q, \forall s \in M_0\},
\end{align*}
\]

then \( \Delta_0(M_0), \Delta_1(M_0) \) and \( \Delta_2(M_0) \) are the stable region, unstable region and the insertion region of \( C_0 \), respectively. Next, we let

\[
g(M_0) = ||\Delta_0(M_0)||, \quad d(M_0) = ||\Delta_1(M_0)||
\]

be the lengths of the stable region and unstable region, respectively, for the partial alignment \( C_0 \).

3. In the stable region \( \Delta_0(M_0) \) and insertion region \( \Delta_2(M_0) \),

\[
\begin{align*}
H_0(M_0) &= \{(j, c_j), j \in \Delta_0(M_0)\}, \quad c_j \neq q, \\
H_2(M_0) &= \{(j, c_j), j \in \Delta_2(M_0)\}, \quad c_j = q
\end{align*}
\]

are the modulus structures of the stable region and insertion region, respectively.

4. In the unstable region \( \Delta_1(M_0) \),

\[
H_1(M_0) = \{(j, c_{M_0,j}), j \in \Delta_1(M_0)\}
\]

is the modulus structure of the unstable region, where \( c_{M_0,j} = \{c_{s,j}, s \in M_0\} \).

These parameters reflect the data structure characteristics of mutation generated by multiple sequence alignments in different aspects. We can alternatively describe these structure characteristics using the network language. Let

\[
\begin{align*}
\tilde{\Delta} &= \{(\Delta_0(M_0), \Delta_1(M_0), \Delta_2(M_0)) : M_0 \subset M\}, \\
H &= \{(H_0(M_0), H_1(M_0), H_2(M_0)) : M_0 \subset M\}
\end{align*}
\]

be the modulus structure of MA.
The Topological Tree Generated by MA

Above, we have shown that $\hat{M} = \{M, D\}$ generated by MA is a metric space. Following from the discussion of Sect. 6.1, we can generate different types of trees according to different data structures, as follows.

Minimum distance clustering tree, minimum distance tree, $k$-order tree, average minimum distance clustering tree, average minimum distance binary tree, average minimum distance binary colored arcs phylogenetic tree. The details of these trees can be found in [35].

The Phylogenetic Tree Generated by a Stable Region and an Unstable Region

In a phylogenetic tree $T' = \{M', V'\}$, let $e = 2m - 1$ be its root, let $T_t = \{M'_t, V'_t\}$ be the branch with root $t(m < t \leq 2m - 1)$, and let $w(e, t')$ be the sum of the lengths of all arcs from $e$ to $t$. $T'$ is then called the phylogenetic tree generated by a stable region and an unstable region of a multiple sequence if $w(e, t') = ||\Delta_0(M_t)|| + ||\Delta_2(M_t)||$, where $M_t$ is the set of all leaves in $T_t$, and $t'$ is the dual point of $t$.

For the phylogenetic trees generated by a stable region and an unstable region of multiple sequences, some properties can easily be found, namely:

1. For any $s \in M$, we always have that $w(e, s') = n$ holds, where $n$ is the length of the MA output $C$.
2. For any $t \in \{m + 1, m + 2, \cdots , 2m - 1\}$ and $s \in M_t$, we always have that $w(e, t') = ||\Delta_0(M_t)|| + ||\Delta_2(M_t)||$, $w(t', s') = ||\Delta_1(M_t)||$

hold, where $\Delta_0(M_t)$ and $\Delta_1(M_t)$ are, respectively, the stable region and unstable region of multiple sequences $M_t$.

3. $w(e, e') = ||\Delta_0(M)||$ is the total length of the common stable region of the MA output. Let $t_1, t_2$ be the two successors of node $e$, the two branches generated by $t_1, t_2$ be $T_{t_1}, T_{t_2}$, and $M_{t_1}, M_{t_2}$ be the sets of leaf nodes of $T_{t_1}, T_{t_2}$. The length of the arcs is then given as

$$
\begin{align*}
\begin{cases}
 w(t_1, t'_1) = ||\Delta_0(M_{t_1})|| + ||\Delta_2(M_{t_1})|| - w(e, e') \\
 = ||\Delta_0(M_{t_1})|| - ||\Delta_0(M)||,
\end{cases}
\end{align*}
$$

$$
\begin{align*}
\begin{cases}
 w(t_2, t'_2) = ||\Delta_0(M_{t_2})|| + ||\Delta_2(M_{t_2})|| - w(e, e') \\
 = ||\Delta_0(M_{t_2})|| - ||\Delta_0(M)||.
\end{cases}
\end{align*}
$$

Similarly, we get the lengths of arcs $w(t, t')$ of all $t \in \{m + 1, m + 2, \cdots , 2m - 1\}$ in the phylogenetic tree $T'$.

4. If $s_1, s_2 \in M$ are the two leaf nodes on the phylogenetic tree $T'$, and they have the same ancestor, then their arc lengths are the penalty function of the alignment sequences $C_{s_1}, C_{s_2}$. That is,

$$
\begin{align*}
w(s_1, s'_1) = w(s_2, s'_2) = ||\Delta_1(s_1, s_2)|| = d(C_{s_1}, C_{s_2}).
\end{align*}
$$
5. The triplet \( T'(w) = \{M', T', w\} \) is called the colored arc graph of the phylogenetic tree \( T' \), where \( w(t, t') \) or \( w(s, s') \) are defined as in (6.33) or (6.32).

6. In the colored arc graph \( T'(w) \) of the phylogenetic tree \( T' \), if we use the stable region and unstable region \( \Delta_0(M_t), \Delta_1(M_s) \) or the modulus structure of the stable region and unstable region \( H_0(M_t), H_1(M_s) \) to replace \( w(t, t') \) and \( w(s, s') \), this colored arc graph turns to the following two forms:

- The colored arc graph of the stable region and unstable region is \( T'(\Delta) = \{M', T', \Delta\} \) if we use \( \Delta_0(t, t') \) or \( \Delta_1(s, s') \) defined as in (6.31) or (6.30).
- The colored arc graph of the stable region and unstable region is \( T'(\mathcal{H}) = \{M', T', \mathcal{H}\} \) if we use the modulus structure \( \mathcal{H} \) defined by (6.31), and \( H_0(t, t') \) or \( H_1(s, s') \) is defined by (6.29) or (6.28).

**Minimum Unstable Region Phylogenetic Tree**

In the above section, we have given the phylogenetic tree \( T' \) generated by the stable region and unstable region. It is simply called the phylogenetic tree \( T' \) of the stable region and unstable region. Let \( w(T') = \sum_{s \in M} w(s, s') \); then it is the sum length of the unstable region of the phylogenetic tree \( T' \).

**Definition 30.** \( T'_0 \) is called the minimum unstable region phylogenetic tree, if \( w(T'_0) \leq w(T') \) holds for all other phylogenetic trees \( T' \).

The method of producing a minimum unstable region phylogenetic tree is similar to that for generating the minimum distance clustering tree. It can be clustered based on the length of the unstable region of the MA output. We will show this later with examples.

**6.2.2 Network System Generated by Mutations of Multiple Sequences**

Among the various topological trees generated from MA outputs, we use graphs and trees to express the connections between mutations and evolution. The modulus structure of the colored arc graph of the stable region and unstable region \( T'(\mathcal{H}) = \{M', T', \mathcal{H}\} \) reflects the information of the MA output. However, some points are less clear for the description of these trees. For example, the combination relations of different sequences within the mutation region are still too complicated to be immediately understood. Therefore, we discuss them further. A network system generated by the mutations of multiple sequences is used to describe the mutation structure of the MA output through the colored arcs graph. To do this, we introduce the following notations.
Network System of Mutation

Let $M = \{1, 2, \cdots, m\}$ be the subscript set of a MA output $C$, that is, each $i \in M$ corresponds to a sequence $C_i$. Then, graphs $G = \{M, V\}$ and $G' = \{M', V'\}$ are generated by MA output $C$, in which $V$ is the arc set generated by the point pairs of $M$, and $\{M', V'\}$ is the extension of $\{M, V\}$ similar to that given by phylogenetic tree $T' = \{M', V'\}$. The network system generated by the MA output colors both points and arcs of the graph $G$ or $G'$. Following from the metric relation $w$ of MA output, two types of network structures may be generated respectively by the mutation region $\Delta$ and the modulus structure $\mathcal{H}$ as follows:

1. Topological network system generated by MA output: $G(w) = \{M, V, W\}$, in which $w$ is the penalty function of the MA output defined by (6.33).
2. Mutation region network system: $G(\Delta) = \{M, V, \Delta\}$, in which $\Delta$ is the mutation region function of the multiple alignment output defined by (6.31).
3. Network system of mutation mode generated by multiple alignment output: $G(\mathcal{H}) = \{M, V, \mathcal{H}\}$, in which $\mathcal{H}$ is the modular function of the multi-sequence alignment given by (6.31).

These three network systems are called the network systems generated by the MA output, or simply the mutation networks. In the same way, we can define the graph $G'$. The purpose in researching the mutation network is to analyze the evolution relations of multiple sequences.

The Basic Mutation Types of Triple Sequences

**Definition 31.** Let $C_1, C_2, C_3$ be a triple sequence in the MA output $C$. Its basic types are stated as follows:

1. **Orthogonal:** Let $\delta_{12}$ and $\delta_{23}$ be the mutation regions induced in $C_1, C_2, C_3$. Then, $H_{12}$ and $H_{23}$ are orthogonal if $\delta_{12} \cap \delta_{23} = \emptyset$. We use the simpler form $H_{12} \perp H_{23}$ to represent the orthogonal relationship.
2. **Overlapping:** The triple sequences $C_1, C_2, C_3$ overlap if their mutations regions satisfy the following: $\delta_{12} = \delta_{13} = \delta_{23}$ and $c_{1j}, c_{2j}, c_{3j}$ are different from each other for all $j \in \delta_{12}$.

**Theorem 26.** 1. The orthogonal type is symmetric. Namely, if $H_{12} \perp H_{23}$ holds, then both $H_{23} \perp H_{12}$ and $H_{21} \perp H_{23}$ hold.
2. $H_{12} \perp H_{23}$ holds if and only if $\delta_{12} \cap \delta_{23} = \emptyset$, in which $\delta_{ij}$ is the mutation region of $H_{ij}$.
3. If $C_1, C_2, C_3$ are overlapping, then $w_{12} = w_{13} = w_{23}$ holds.

It is easy to prove these three propositions, so we omit the proofs here.

The orthogonal type and overlapping type are the two extreme cases for mutations. In general, we frequently face mixed modes. Thus, we need the following decomposition theorem:
Theorem 27. (The decomposition theorem of the triple alignment output.)
Let $C_1, C_2, C_3$ be the alignment output of the triple sequence $A_1, A_2, A_3$. There is a new triple sequence $C'_1, C'_2, C'_3$ satisfying the following properties:

1. $C'_1, C'_2, C'_3$ are overlapping.
2. Mutation modes $H'_{11}, H'_{22}, H'_{33}$ are orthogonal to each other. As well, $H'_{11'}$ and $H'_{1'2}, H'_{1'3}, H'_{22'}$ and $H'_{2'1}, H'_{2'3}, H'_{33'}$ and $H'_{3'1}, H'_{3'2}$ are all orthogonal.

Remark 6. 1. In Fig. 6.5a, $C_1, C_2, C_3$ represent the alignment output of the triple sequence $A_1, A_2, A_3$, where $N - \delta$ is the stable region, in which the values of these three sequences are the same. $\delta$ is the unstable region, which can be decomposed to four subregions $\delta_1, \delta_2, \delta_3$, and $\delta_0$ as shown in (6.34).
2. Figure 6.5b shows sequences $C'_1, C'_2, C'_3$ defined by (6.35).

Proof. Maintaining the notation given in Fig. 6.5, $C_1, C_2, C_3$ are the alignment output of triple sequences $A_1, A_2, A_3$ and the mutation region is $\delta$. $N' - \delta$ is then the stable region. The unstable region can be decomposed into $\delta_{12}, \delta_{13}, \delta_{23}$. 

Fig. 6.5a,b. The decomposition of the mutation region of a triple alignment output.
\[ \delta_{23}, \delta_{13}. \] These are the mutation regions of \((C_1, C_2), (C_2, C_3), (C_1, C_3). \] Let

\[
\begin{align*}
\delta &= \delta_{12} \cup \delta_{13} \cup \delta_{23}, \\
\delta_0 &= \{ j \in \delta : c_{1j}, c_{2j}, c_{3j} \text{ are not the same as each other} \}, \\
\delta_1 &= \{ j \in \delta : c_{1j} = c_{2j} \neq c_{3j} \}, \\
\delta_2 &= \{ j \in \delta : c_{1j} = c_{3j} \neq c_{2j} \}, \\
\delta_3 &= \{ j \in \delta : c_{2j} = c_{3j} \neq c_{1j} \},
\end{align*}
\]

then \(\delta_0, \ldots, \delta_3\) are four mutually disjoint regions. We denote the lengths of the four regions by \(w_\tau = ||\delta_\tau||\), where \(\tau = 0, 1, 2, 3\), respectively. Based on this decomposition, we construct new sequences \(C_1', C_2', C_3'\) as follows. Let

\[
\begin{align*}
c_{1'j} &= \begin{cases} c_{2j}, & \text{if } j \in \delta_3, \\ c_{1j}, & \text{otherwise}, \end{cases} & c_{2'j} &= \begin{cases} c_{1j}, & \text{if } j \in \delta_2, \\ c_{2j}, & \text{otherwise}, \end{cases} \\
c_{3'j} &= \begin{cases} c_{1j}, & \text{if } j \in \delta_1, \\ c_{3j}, & \text{otherwise}. \end{cases}
\end{align*}
\]

Then, the components of sequences \(C_1', C_2', C_3'\) are different from each other in the region \(\delta_0\) but the same in the remaining regions. Therefore, it is the overlapping type. In addition, we analyze the mutation regions of sequences \(C_1, C_2, C_3\) and \(C_1', C_2', C_3'\) as follows. Since we then have

\[ \delta_{11'} = \delta_3, \quad \delta_{2'2} = \delta_2, \quad \delta_{33'} = \delta_1, \]

and since regions \(\delta_1, \delta_2, \delta_3\) are mutually disjoint, it follows that \(\{H_{11'}, H_{22'}, H_{33'}\}\) are orthogonal modulus structure. With the same reasoning, we may prove that the three groups of modes \(H_{11'}, H_{1'2}, H_{1'3}; H_{22'}, H_{2'1}, H_{2'3}; H_{33'}, H_{3'1}, H_{3'2}\) are orthogonal, respectively. Thus ends the proof.

Figure 6.6 shows the mutation relations between sequences \(C_1, C_2, C_3\) and \(C_1', C_2', C_3'\). The process by which \(C_1\) mutates to \(C_2, C_3\) can be decomposed, to where \(C_1\) mutates to \(C_1'\), and then \(C_1'\) mutates to \(C_2, C_3\). Therefore, Fig. 6.6 is called the network structure graph of the triple alignment output.

In Theorem 27, the triangle \(\Delta(C_1', C_2', C_3')\) shrinks to a point if \(\delta_0\) is an empty set. If \(C_1' = C_2' = C_3' = C_0\) are the same sequences, then

\[ H_{10} \perp H_{20}, \quad H_{10} \perp H_{30}, \quad H_{20} \perp H_{30} \quad (6.36) \]

hold. The inverse proposition is also true, i.e., if there is a point \(C_0\) making (6.36) true, then \(\delta_0\) must be an empty set.

**Definition 32.** 1. Under the conditions of Theorem 27, \(C_1', C_2', C_3'\) are the orthogonal decomposition of the triple alignment output \((C_1, C_2, C_3)\) if sequences \(C_1', C_2', C_3'\) satisfy the theorem.
2. If there is a sequence $C_0$ such that (6.36) holds, then we say that $C_0$ makes the triangle $(C_1, C_2, C_3)$ perfectly orthogonal. The triple sequences $(C_1, C_2, C_3)$ can be made perfectly orthogonal if and only if $C_1' = C_2' = C_3'$ holds, where the mutation relationship of $C_1, C_2, C_3$ can be decomposed to the mutation relationship between $C_1, C_2, C_3$ and $C_1', C_2', C_3'$.

The Mutation Network Tree Generated by a Binary Tree

In any book on graph theory, the reader can find the terms graph, tree, directed tree, node, arc, the extreme points of an arc, the starting point and end point of the directed arc, the root of a tree, and leaf all well-defined. Therefore, we do not repeat the definitions here. However, several new concepts are directly involved in the discussions presented in this book, which are defined as follows:

**Definition 33.** 1. For a mutation network $\mathcal{E}$, if each overlapping triangle is seen as a point, then the renewed mutation network $\mathcal{E}'$ is the reduction of $\mathcal{E}$.
2. A directed mutation network tree is a directed orthogonal mutation tree if any two arcs starting from any node are orthogonal.
3. An undirected mutation network tree is a perfectly orthogonal mutation tree if any two arcs with a common node are orthogonal.

**Theorem 28.** (The orthogonalization theorem of a mutation network tree.) For a given directed mutation network tree, there are some nodes such that the mutation network $\mathcal{E}$, which is generated by adding these nodes into the given tree, satisfies the following conditions:

1. If there are triangles in $\mathcal{E}$, they are overlapping triangles.
2. Let $\mathcal{E}'$ be the network induced by $\mathcal{E}$ in the case where each overlapping triangle is seen as a point, then $\mathcal{E}'$ is an orthogonal mutation tree.
3. Let $\Delta(a,b,c)$ be an overlapping triangle in the mutation network $E$. Each arc with an extreme point $a$ is then orthogonal to arcs $ab, ac$. Also, the same holds true for both $b$ and $c$.

**Proof.** For clarity, we follow Fig. 6.7 to give the proof as follows:

1. Figure 6.7a is the original undirected tree, where $G_1^{(0)} = \{M^{(0)}, V^{(0)}\}$ and $M^{(0)} = \{a, b, c, d, e\}$, $V^{(0)} = \{(a, b), (a, c), (b, d), (b, e)\}$. The virtual lines are 2-order arcs.

2. The orthogonalization starts from leaves $a, c$. Following from Theorem 27, there is an overlapping triangle $\Delta(a', b', c')$ which orthogonalizes $(a, b, c)$. The modes $H_{aa'}, H_{bb'}, H_{cc'}$ are orthogonal to each other. If we reduce the network graph such that $ab, ac, bc$ are seen as 2-order arcs, then we get Fig. 6.7b, and its mutation network tree is $G_1^{(1)} = \{M^{(1)}, V^{(1)}\}$, where

$$
\begin{align*}
M^{(1)} &= \{a, b, c, d, e, a', b', c'\}, \\
V^{(1)} &= \{(a, a'), (b, b'), (c, c'), (a', b'), (a', c'), (b', c'), (b, d), (b, e)\}.
\end{align*}
$$

Fig. 6.7a–d. The orthogonalization procedure of a mutation network
3. Similarly to step 2, we orthogonalize triangle \((b, b', e)\) in \(G_1^{(1)}\). If this triangle is perfectly orthogonal, then Fig. 6.7c is obtained. Its mutation network tree is then \(G_1^{(2)} = \{M^{(2)}, V^{(2)}\}\), where

\[
\begin{align*}
M^{(2)} &= \{a, b, c, d, e, a', b', c', f\}, \\
V^{(2)} &= \{(a, a'), (b, b'), (c, c'), (a', b'), (a', c'), (b', c'), (b', f), (b, f), \\
& \quad (f, e), (b, d)\}.
\end{align*}
\]

4. Continuing this procedure, we can do the orthogonalization procedure on \(G_1^{(2)}\). Finally, we get \(G_1^{(3)}\) as shown by Fig. 6.6d, where

\[
\begin{align*}
M^{(3)} &= \{a, b, c, d, e, a', b', c', f, f', b'', d'\}, \\
V^{(3)} &= \{(a, a'), (b, b'), (c, c'), (a', b'), (a', c'), (b', c'), (b', f), (f, f'), \\
& \quad (b'', f'), (f', d'), (b'', d'), (d', d), (b, b''), (f, e)\}.
\end{align*}
\]

The graph \(G_1^{(3)}\) satisfies all the conditions in the theorem. Thus ends the proof.

We have introduced the mutation network of a MA output, with the intention that we may easily obtain the mutation relations of data structures among a multiple sequence by viewing these graphs. For example, viewing Fig. 6.7d, we find that the mutation process from sequence \(a\) to \(d, e\) can be decomposed as follows:

\[
\begin{align*}
a &\rightarrow a' \rightarrow b' \rightarrow f \rightarrow e, \\
&\rightarrow a' \rightarrow b' \rightarrow f \rightarrow f' \rightarrow d' \rightarrow d,
\end{align*}
\]

in which, \(a \rightarrow a' \rightarrow b' \rightarrow f \rightarrow e\) are perfectly the same type, and \(f \rightarrow e\) and \(f \rightarrow f' \rightarrow d'\) are mutually orthogonal. Typically, the mutation process of each smaller segment is orthogonal. Following from this, we can deduce the relations of the mutation network of any multiple sequences.

### 6.3 The Application of Mutation Network Analysis

MA and the application of mutation network analysis can be used in many fields of biological research. We discuss the evolution and development of epidemics in the following.

#### 6.3.1 Selection of the Data Sample

To examine the evolution of biosomes on a molecular level, we should begin with the proper selection of data. We always use DNA, RNA, or protein databases. The requirement for the use of these databases is that a sequence should have many homologous sequences in different biosomes. Research in
biology indicates that many genes and proteins recur in many species. For example, chondriosome, cytochrome and cathepsin are found in many biosomes. In the process of selecting data samples, besides using existing data that may be obtained directly from databases such as GenBank, some special databases may also need to be tracked. Therefore, we need to design the data collection scheme before starting the sequencing. For example, to analyze the development of some epidemic or disease, we must design a good scheme for collecting the required data. Next, we choose chondriosome, SARS, and HIV-1, respectively, as examples to illustrate the procedure used to analyze the data. The explanation for the corresponding results is given below:

**The Data Sample of Chondriosome**

Biology research has revealed that chondriosome occurs in many biosomes. In GenBank, there are thousands of homologous sequences of chondriosome. To analyze the mutations, we select the ND1 gene coding region of 20 species of mammals as follows:

1. Bos taurus complete mitochondrial
2. Balaenoptera physalus mitochondrial, complete
3. Balaenoptera musculus mitochondrial DNA, complete
4. Phoca vitulina mitochondrial DNA, complete
5. Halichoerus grypus complete mitochondrial
6. Felis catus mitochondrion, complete
7. Equus caballus mitochondrial DNA, complete sequence
8. Rhinoceros unicornis complete mitochondrial
9. Rattus norvegicus mitochondrial
10. Homo sapiens mitochondrial DNA, complete sequence
11. Pan troglodytes mitochondrial DNA, complete sequence
12. Pan paniscus mitochondrial DNA, complete sequence
13. Gorilla gorilla mitochondrial DNA, complete sequence
14. Pongo pygmaeus mitochondrial DNA, complete sequence
15. Pongo pygmaeus abelii mitochondrial
16. Hylobates lar complete mitochondrial DNA sequence
17. Didelphis virginiana complete mitochondrial
18. Macropus robustus complete mitochondrial
19. Ornithorhynchus anatinus mitochondrial DNA, complete
20. Mus musculus mitochondrial

**SARS Sequences**

In the spring of 2003, a SARS epidemic broke out in China. Research on the SARS virus has become an important problem in the fields of biology and medicine. In the GenBank database, new DNA sequences of SARS were continually announced. In September of 2003, an article published in *Science* [101]
involved 63 DNA sequences of the SARS virus. As a result, this paper analyzed the evolution of the SARS epidemic from its onset to the metaphase and then to the mature phase. After September 2003, more new DNA sequences of the SARS virus were constantly being sequenced. As of September 2004, the total number of SARS virus sequences uploaded in the GenBank was 118. Their names and sources are shown in Table 6.1.

Remark 7. Under the “name” rubric, we only give the simpler name of the SARS coronavirus. For example, in number 4, we only use the name Sin850, while its full name is SARS coronavirus Sin850. Pagumalarvata is the Chinese southern Pagumalarvata. The CDC is CDC-200301157, Pagu. is Pagumalarvata, SH stands for Shanghai.

HIV-1 Virus Genome

The HIV-1 virus genome is the main type of AIDS virus. Besides HIV-1, there is HIV-2 along with other virus genomes of animals. Since HIV-2 appears in local districts, most studies of the AIDS virus genome focus on how to analyze the HIV-1 virus genome. In edition 2004/9 (release 43), the GenBank announced 706 sequences of HIV-1. The lengths of these sequences vary from 7,000 to 9,000bp. Similarly to the SARS sequences, HIV-1 data contain both incomplete regions and nonsequenced regions. Therefore, we cannot adopt them mechanically. The nations and districts of origin for these 706 sequences of HIV-I are listed in Table 6.2.

6.3.2 The Basic Steps to Analyze the Sequences

The data samples we collected are a group of multiple sequences. Therefore, we process them by using various types of software packages to obtain a MA output. Let $\mathcal{A}$ be the multiple sequences consisting of the data samples, and let $\mathcal{A}'$ be its MA output.

The Procedure to Analyze the MA Output

1. Based on $\mathcal{A}'$, we compute the penalty (or scoring) matrix $W = (w_{s,t})$, modulus structure matrix $H = (H_{s,t})$ and mutation region matrix $\Delta$. Because the modulus structure matrix and the mutation region matrix $H$, $\Delta$ are very complex, they may be considered to be parameters.
2. Based on the penalty matrix $W = (w_{s,t})$ to cluster the multiple sequences, we construct the minimum distance tree $G_1$, and then construct the $k$-order graph $G_k$ and $k$-order mutation network $G_k(W)$.
3. Based on the minimum distance tree $G_1$ and mutation region matrix $\Delta$, we orthogonalize the network, and give the corresponding graph for the orthogonal decomposition of the network.
### Table 6.1. The names and numbered list of the 118 SARS sequences

| No. | GenBank number | Name       | Nation or district | No.  | GenBank number | Name       | Nation or district |
|-----|----------------|------------|--------------------|-----|----------------|------------|--------------------|
| 1   | NC-004718      | Toronto    |                    | 2   | AY714217       | CDC        | USA                |
| 3   | AY559097       | Sin3408L   | Singapore          | 4   | AY559096       | Sin850     | Singapore          |
| 5   | AY559095       | Sin847     | Singapore          | 6   | AY559094       | Sin846     | Singapore          |
| 7   | AY559093       | Sin845     | Singapore          | 8   | AY559092       | SinP5      | Singapore          |
| 11  | AY559089       | SinP2      | Singapore          | 12  | AY559088       | SinP1      | Singapore          |
| 13  | AY559087       | Sin3725V   | Singapore          | 14  | AY559086       | Sin849     | Singapore          |
| 15  | AY559085       | Sin848     | Singapore          | 16  | AY559084       | Sin3765V   | Singapore          |
| 17  | AY559082       | Sin852     | Singapore          | 18  | AY559081       | Sin842     | Singapore          |
| 19  | AY654624       | TJF        | Beijing            | 20  | AY595412       | LLJ-2004   | Beijing            |
| 21  | AY394850       | WHU        | Wuhan              | 22  | AY274119       | Tor2       | Toronto            |
| 23  | AY323977       | HSR 1      | Italy              | 24  | AY291315       | Frankfurt1 | Germany            |
| 25  | AY502932       | TW9        | Taiwan             | 26  | AY502931       | TW8        | Taiwan             |
| 27  | AY502930       | TW7        | Taiwan             | 28  | AY502929       | TW6        | Taiwan             |
| 29  | AY502928       | TW5        | Taiwan             | 30  | AY502927       | TW4        | Taiwan             |
| 31  | AY502926       | TW3        | Taiwan             | 32  | AY502925       | TW2        | Taiwan             |
| 33  | AY502924       | TW11       | Taiwan             | 34  | AY502923       | TW10       | Taiwan             |
| 35  | AY291451       | TW1        | Taiwan             | 36  | AY390556       | GZ20       | Guangdong          |
| 37  | AY395003       | ZS-C       | Guangdong          | 38  | AY395002       | LC5        | Guangdong          |
| 39  | AY395001       | LC4        | Guangdong          | 40  | AY395000       | LC3        | Guangdong          |
| 41  | AY394999       | LC2        | Guangdong          | 42  | AY394998       | LC1        | Guangdong          |
| 43  | AY394997       | ZS-A       | Guangdong          | 44  | AY394996       | ZS-B       | Guangdong          |
| 45  | AY394995       | HSZ-Cc      | Guangdong          | 46  | AY394994       | HSZ-Bc      | Guangdong          |
| 47  | AY394993       | HGZ6L2      | Guangdong          | 48  | AY394992       | HSZ2-C      | Guangdong          |
| 49  | AY394991       | HSZ2-Fc     | Guangdong          | 50  | AY394990       | HSZ2-E      | Guangdong          |
| 51  | AY394899       | HSZ2-D      | Guangdong          | 52  | AY39487       | HSZ2-Fb     | Guangdong          |
| 53  | AY394896       | HSZ2-Cb     | Guangdong          | 54  | AY39485       | HSZ2-Bb     | Guangdong          |
| 55  | AY394893       | HSZ2-A      | Guangdong          | 56  | AY39482       | HGZ8L1-B    | Guangdong          |
| 57  | AY394891       | HGZ8L1-A    | Guangdong          | 58  | AY39479       | GZ-C        | Guangdong          |
| 59  | AY394798       | GZ-B        | Guangdong          | 60  | AY508724      | NS-1       | Guangdong          |
| 61  | AY463059       | SH-QXC1     | Guangdong          | 62  | AY313906      | GD69       | Guangdong          |
| 63  | AY310120       | FRA         | Italy              | 64  | AY461660      | SoD        | Russia             |
| 65  | AY485278       | Sino3-11    | Beijing            | 66  | AY485277      | Sino1-11   | Beijing            |
| 67  | AY345988       | CUHK-AG03   | Hong Kong          | 68  | AY345987      | CUHK-AG02   | Hong Kong          |
| 69  | AY345986       | CUHK-AG01   | Hong Kong          | 70  | AY282752      | CUHK-Su10   | Hong Kong          |
| 71  | AY357076       | PUMC03      | Beijing            | 72  | AY357075      | PUMC02     | Beijing            |
| 73  | AY350750       | PUMC01      | Beijing            | 74  | AY304495      | GZ50       | Hong Kong          |
| 75  | AY304486       | SZ3         | Pagu               | 76  | AY427439      | AS         | Italy              |
| 77  | AY283798       | Sin2774     | Singapore          | 78  | AY278491      | HUK-39849  | Beijing            |
| 79  | AY278489       | GD01        | Beijing            | 80  | AY362699      | TWC3       | Taiwan             |
| 81  | AY362628       | TWC2        | Taiwan             | 82  | AY283797      | Sin2748    | Singapore          |
| 83  | AY283796       | Sin2679     | Singapore          | 84  | AY283795      | Sin2677    | Singapore          |
| 84  | AY283794       | Sin2500     | Singapore          | 85  | AY287841      | Urbani     | USA                |
| 86  | AY351680       | ZMY 1       | Guangdong          | 88  | AP006561      | TWY        | Taiwan             |
| 89  | AP006560       | TWS         | Taiwan             | 90  | AP006559      | TWK        | Taiwan             |
| 91  | AP006558       | TWJ         | Taiwan             | 92  | AP006557      | TWK        | Taiwan             |
| 93  | AY278554       | CUHK-W1     | Hong Kong          | 94  | AY348314      | TaiwanTC3  | Taiwan             |
| 95  | AY338175       | Taiwan TC2  | Taiwan             | 96  | AY338174      | TaiwanTC1  | Taiwan             |
| 97  | AY321118       | TWC         | Taiwan             | 98  | AY279354      | BJ04       | Beijing            |
| 99  | AY278490       | BJ03        | Beijing            | 100 | AY278487      | BJ02       | Beijing            |
| 101 | AY297028       | ZJ01        | Beijing            | 102 | AY278488      | BJ01       | Beijing            |
| 103 | AY304488       | SZ16        | Pagu               | 104 | AY559083      | Sin3408    | Shanghai           |
| 105 | AY286320       | ZJ01        | Hangzhou           | 106 | AY395004      | HZS2-Bb    | Guangdong          |
| 107 | AY394898       | JMD         | Guangdong          | 108 | AY394894      | SHZ-A       | Guangdong          |
| 109 | AY394890       | GZ-D        | Guangdong          | 110 | AY394977      | GZ-A        | Guangdong          |
| 111 | AY463060       | SH-QXC2     | Shanghai           | 112 | AY304494      | HUK-66078  | Hong Kong          |
| 113 | AY304493       | HKU-65806   | Hong Kong          | 114 | AY304492      | HKU-36871  | Hong Kong          |
| 115 | AY304491       | GZ60        | Hong Kong          | 116 | AY304490      | GZ43       | Hong Kong          |
| 117 | AY304489       | SZ1         | Pagu               | 118 | AY304487      | SZ13       | Pagu               |
Table 6.2. The nations and districts for the 706 sequences of HIV-I

| A          | B A | A B | B A | B A | B A | B A | B A |
|------------|-----|-----|-----|-----|-----|-----|-----|
| Botswana   | 72  | Tanzania | 41 | Cameroon | 82 | South Africa | 46 | DR Congo | 11 | Senegal | 7 |
| Ethiopia   | 8   | Nigeria | 3 | Zambia | 2 | Rwanda | 1 | Benin | 1 | Uganda | 58 |
| Kenya      | 45  | Gabon | 2 | Central African | 5 | Chad | 3 | Niger | 3 | Mali | 2 |
| Finland    | 3   | Belgium | 11 | France | 23 | Sweden | 15 | Greece | 4 | Belarus | 2 |
| Russia     | 3   | Spain | 14 | Netherlands | 14 | Estonia | 2 | Britain | 2 | Germany | 2 |
| Ukraine    | 1   | Norway | 1 | Taiwan | 1 | South Korea | 2 | China | 17 | Israel | 1 |
| India      | 15  | Thailand | 59 | Ghana | 3 | Japan | 4 | Myanmar | 9 | Cyprus | 2 |
| Brazil     | 7   | Uruguay | 4 | Argentina | 26 | Bolivia | 2 | Colombia | 5 | Australia | 16 |
| USA        | 43  | Others | 1 |

A denotes the nation or district, B denotes the numbers of the sequenced genes.

Analyzing the Biological Meaning of the Final Results

Based on the graph of the orthogonal decomposition of the network, we can construct a relationship of the mutations among sequences, and analyze the biological meaning. For the same biosome, there are many methods to collect the data sample, and for different data samples we may get different results. Therefore, we should analyze the biological meaning from several different angles.

Using the above general procedure, we next discuss several examples in biology and medicine. We will detail the content involved within the discussions.

6.3.3 Remarks on the Alignment and Output Analysis

The Mutation Analysis of Mammalian Mitochondrial Genome

1. The length of the mammalian mitochondrial genome is about 18 kbp. The length of the coding region ND1 is 900 bp. The length of its alignment output is 961 bp, as shown in [99].
2. The total length of the stable region of the multiple alignment output is $||N_0|| = 404$, the percentage is $404/961 = 42.04\%$, proportional to the total length of the output. While we can readily produce the list of their modulus structure, we have omitted it for brevity.
3. Let $w(a, b)$ be the Hamming matrix, and let the penalty matrix be $w_{s,t} = \sum_{i=1}^{n} w(c_{s,i}, c_{t,i})$, where $s, t = 1, 2, \cdots, 20$ as shown in Table 6.3.
4. Based on the penalty matrix, we find the system clustering tree as shown in Fig. 6.8.

The Analysis of the SARS Virus Gene

1. The lengths of the 118 SARS sequences are about 18 kbp. We select 103 sequences which are well-sequenced. The length of the MA output is 29,908 bp. The result is shown in [99].
6.3 The Application of Mutation Network Analysis

Table 6.3. The penalty matrix for the ND1 coding region

| 180  | 178  | 175  | 29  |
|------|------|------|-----|
| 192  | 191  | 183  | 152 |
| 153  | 148  | 152  | 180 |
| 192  | 191  | 183  | 152 |
| 153  | 148  | 152  | 180 |
| 153  | 148  | 152  | 180 |
| 153  | 148  | 152  | 180 |

2. The SARS virus genome has high similarity because of the short time the disease has taken to develop and evolve. Except for a few sequences which may have sequencing errors, the sequence homology for most sequences is over 95%. In these 103 SARS sequences, we have determined their common stable region (at whose positions the nucleotides are invariant). The number of the positions in the common stable region is 26,924, which is 90.023% of the length of the sequence alignment output (29,908).

3. Analyze the unstable region of the MA output from different angles, including the head and tail of the SARS sequences. For the MA output, we can determine that the head comprised 20 positions and the tail comprised 43 positions. The percentages are 0.07% and 0.144%, respectively. In the head and tail part, the structure changes a great deal. The reason is that the start point and end point are both selected differently in sequencing. The distribution of the nucleotides in unstable positions can be denoted by $\tilde{f}_i = (f_i(0), f_i(1), \ldots, f_i(4))$, where $f_i(z)$ is the number of nucleotides or inserting symbols $z$ at position $i$. For example, $f_{19} = (1, 86, 0, 1, 15)$ means that that number of times that “a, c, g, t, and −” occur at position 19 of the 103 SARS sequences are 1, 86, 0, 1, and 15, respectively.

4. The penalty matrix $W = (w_{s,t})_{s,t=1,2,\ldots,103}$ follows from the multiple alignment output (shown in [99]), where $w_{s,t} = d_H(A_s, A_t)$ is the Hamming distance between $A_s$ and $A_t$.

5. Following from the penalty matrix $W$, we generate the phylogenetic tree, the minimum distance graph and the second-order structure graph. Construction of the network graph follows directly from these graphs.
Fig. 6.8. The cluster tree generated by the multiple sequence alignment of the ND1 gene coding region of 20 sorts of mammals

1. Bos taurus 2. Balaenoptera physalus
3. Balaenoptera musculus 4. Phoca vitulina
5. Halichoerus grypus 6. Felis catus
7. Equus caballus 8. Rhinoceros unicornis
9. Rattus norvegicus 10. Homo sapiens
11. Pan troglodytes 12. Pan paniscus
13. Gorilla gorilla 14. Pongo pygmaeus
15. Pongo pygmaeus abelii 16. Hylobates lar
17. Didelphis virginiana 18. Macropus robustus
19. Ornithorhynchus anatinus 20. Mus musculus

The Network Graph Based on the SARS Sequences in Different Stages

In clinics, a disease is divided into many stages. SARS, as a particular disease, is also divided into an initial stage, a middle stage and a final stage. The SARS sequences change due to mutations as the stage or other conditions change.
6.3 The Application of Mutation Network Analysis

To search the variance, we discuss the network graph based on the sequences collected at different stages. The discussion is detailed as follows:

1. For some sequences, for example, numbers 42, 50, and 51, the differences among them are very minor. They always come from the same district. It is useful to track their evolution processes (e.g., the time point for the onset, the development of the epidemic process, etc.).

2. Some sequences, e.g., numbers 5, 28, 76, and 93, form local clustering centers in the graph. These centers can be seen as sources of SARS in some districts.

3. Sequence 75 is the sequence of Pagumalarvata SZ3 (GenBank number: AY304486) (see [101]), the prevalent conclusion (including the conclusion in [101]) is that Pagumalarvata is the source of the SARS virus. However, based on the structure in Fig. 6.9, this conclusion can be challenged. If sequences 75, 36, and 47 were sequenced correctly, then $75 \rightarrow 36 \rightarrow 47$, and double mutations happened at positions 48 and 68. If this conclusion is right, then the double mutations are the key causes of the SARS outbreak in 2003.

Fig. 6.9. The network graph based on the SARS sequences

[Diagram of the network graph based on SARS sequences showing nodes and edges with labels and connections between sequences 1 to 75 and others like 42, 38, 39, 40, 41, among others.]

To search the variance, we discuss the network graph based on the sequences collected at different stages. The discussion is detailed as follows:

1. For some sequences, for example, numbers 42, 50, and 51, the differences among them are very minor. They always come from the same district. It is useful to track their evolution processes (e.g., the time point for the onset, the development of the epidemic process, etc.).

2. Some sequences, e.g., numbers 5, 28, 76, and 93, form local clustering centers in the graph. These centers can be seen as sources of SARS in some districts.

3. Sequence 75 is the sequence of Pagumalarvata SZ3 (GenBank number: AY304486) (see [101]), the prevalent conclusion (including the conclusion in [101]) is that Pagumalarvata is the source of the SARS virus. However, based on the structure in Fig. 6.9, this conclusion can be challenged. If sequences 75, 36, and 47 were sequenced correctly, then $75 \rightarrow 36 \rightarrow 47$, and double mutations happened at positions 48 and 68. If this conclusion is right, then the double mutations are the key causes of the SARS outbreak in 2003.
**Mutation Network Structure of Early SARS Sequences**

We compare the SZ16 (a) and SZ3 (b) of Pagumalarvata with SARS sequences HSZ-Bc (AY394994), GZ02 (AY390556), HSZ-Cc (AY394995), HSZ-Cb (AY394986) in the early period with the HZS2-E (AY394990) in the metaphase, respectively. We number the seven sequences as SZ16 (a), SZ3 (b), HSZ-Bc (c), GZ02 (d), HZS2-E (e), HSZ-Cb (u), HSZ-Cc (v), respectively. We then analyze their mutation network structure, and obtain the following result:

1. From the MA output of the SARS sequences, we find that

\[
\begin{align*}
    w_{ac} &= 50, \\
    w_{au} &= 87, \\
    w_{cu} &= 37, \\
    w_{cv} &= 6, \\
    w_{av} &= 56, \\
    w_{ac} + w_{cu} &= w_{au}, \\
    w_{ac} + w_{cv} &= w_{av}.
\end{align*}
\]

This implies that arc ac is orthogonal to cu, cv. We conclude that the SARS virus starts from SZ16 (a) (Pagumalarvata), to HSZ-Bc (c), then from HSZ-Bc to HSZ-Cb (u) and HSZ-Cc (v), respectively. i.e., the source of HSZ-Bc is SZ16, while the cause of both HSZ-Cb and HSZ-Cc is SZ16.

2. In the infection process where the SARS virus progresses from SZ16 to HSZ-Bc and then to HSZ-Cb and HSZ-Cc, the number of times mutations occur is 50, 37, and 3, respectively, and the mutation modes are also determined by the MA output.

3. The source of HSZ-Cb (u) and HSZ-Cc (v) is determined, we need only discuss the mutation structure of SZ16 (a), SZ3 (b), HSZ-Bc (c), GZ02 (d) and HZS2-E (e). This discussion is given below.

**Remark 8.** 1. Points a, b, c, d, e represent the five SARS sequences in the initial stage. Points f, g, h are the transitional sequences in orthogonal decomposition.

2. In the distance graph constructed by a, b, c, d, e, f, g, h nodes; thick lines are first-order arcs, and thin straight lines are second-order arcs. The numbers written on the sides of the lines represent the mutation errors.

**The Analysis of the Network Structure Graph – Fig. 6.10**

1. The triangles in the network structure, first-order arcs and second-order arcs are orthogonal. For example, in triangle \(\Delta(a, b, h)\), the formula

\[
|ab| = |ag| + |bg|, \quad |ah| = |ag| + |gh|, \quad |bh| = |bg| + |gh|
\]

holds.

2. For the 1st-order arcs in Fig. 6.10, the modulus structures are orthogonal to each other. For example, \(H_{ag}, H_{bg}, H_{hg}\) are mutually orthogonal.

3. In the SARS virus genome of Pagumalarvata, there are 20 mutation differences, and the mutation mode is \(H_{ab}\). It may be decomposed orthogonally as \(H_{ab} = H_{ag} + H_{bg}\).
6.3 The Application of Mutation Network Analysis

Fig. 6.10. The mutation network decomposition of SARS sequences in the initial stage

Table 6.4. The structural representation of the mutation mode $H_{gh}$

| Mutation gh position | Mutation gh position | Mutation gh position | Mutation gh position | Mutation gh position |
|----------------------|----------------------|----------------------|----------------------|----------------------|
| 1899                 | TG                   | 3664                 | TC                   | 6455                 | GA                   | 69                    | CT                   | 13882                | TC                   |
| 22216                | AC                   | 22317                | AC                   | 22615                | CT                   | 22974                | AT                   | 22997                | GC                   |
| 23356                | CT                   | 23531                | CT                   | 23641                | TC                   | 23768                | GA                   | 23802                | TC                   |
| 24221                | GA                   | 25340                | AT                   | 25562                | AT                   | 25598                | TC                   | 25682                | GT                   |
| 26464                | AG                   |                      |                      |                      |                      |                      |                      |                      |                      |

4. When the SARS virus of Pagumalarvata infects human beings, the mutations of the genome consist of three parts: the first part is the mutation differences (i.e., $H_{ag}, H_{bg}$) of different Pagumalarvatas; the second part is the mutation differences (i.e., $H_{hf}, H_{hc}$) of different human beings; and the third part is the common mutation differences (i.e., $H_{gh}$) of human beings and Pagumalarvatas. We believe that the particular mutation $H_{gh}$ is the key to how Pagumalarvata infects human beings. The mutation mode is shown in Table 6.4.

Remark 9. The mutation position in the table is where the mutation happens. The capital letters are the nucleotides which mutate, e.g., 1899 TG means that the nucleotides in sequences g and h at the 1899th position of the alignment output are T and G, respectively.

5. After the SARS virus of Pagumalarvata infected human beings, many cases emerged. However, the SARS disease may break out only if the HZS2-E(e) virus occurs. Therefore, the mutation $H_{fe}$ is the key to a SARS outbreak. The mutation modes are as shown in Table 6.5.

Remark 10. The data in Table 6.5 are defined the same way as those in Table 6.4.
Table 6.5. The structural representation of the mutation mode $H_{fe}$

| Mutation fe position | Mutation fe position | Mutation fe position | Mutation fe position | Mutation fe position |
|----------------------|----------------------|----------------------|----------------------|----------------------|
| 1196 CT              | 9406 CT              | 9481 CT              | 14606 CT              | 20884 AG             |
| 23873 GT             | 25028 GA             | 27945 CA             | 27946 C−              | 27947 T−             |
| 27948 A−             | 27949 C−             | 27950 T−             | 27951 G−              | 27952 G−             |
| 27953 T−             | 27954 T−             | 27955 A−             | 27956 C−              | 27957 C−             |
| 27958 A−             | 27959 A−             | 27960 C−             | 27961 C−              | 27962 T−             |
| 27963 G−             | 27964 A−             | 27965 A−             | 27966 T−              | 27967 G−             |
| 27968 G−             | 27969 A−             | 27970 A−             | 27971 T−              | 27972 A−             |
| 27973 T−             | 27974 A−             |                      |                      |                      |

Remark 11. The results listed in Tables 6.4 and 6.5 are mathematical results. They may be used as a reference for biology and medicine. Whether or not these results are correct must still be proved through observations and experiments.

The Alignment Output of the Sequences of HIV-1

Amongst the 706 HIV-1 sequences, we select 704 better sequences to be aligned. The lengths of the 704 HIV-1 sequences are within 7000–9000bp. We produce the alignment output which is a $704 \times 11,364$ matrix. Because the 704 HIV-1 sequences refer to many nations or districts over a long time, we omit discussion of the alignment output.

6.4 Exercises, Analyses, and Computation

Exercise 29. Construct the phylogenetic tree and graph based on the penalty matrix in Sect. 6.3.3, according to the requirements listed below:

1. Minimum distance phylogenetic clustering tree, and the average minimum distance phylogenetic clustering tree
2. Directed and undirected minimum distance tree
3. Minimum distance two-order tree

Exercise 30. The ND1 gene coding region sequences of 20 species of mammals, and the MA outputs for 103 SARS sequences and 706 HIV-1 sequences are included on our Web site [99]. Construct the mutation network based on these datasets. Compute the stable and unstable regions for them, and represent them using modulus structure.

Exercise 31. Compute the similarity matrices of the MA outputs of the SARS sequences and HIV-1 sequences, and analyze the phylogenetic tree based on them. Also compute the following results:
1. Construct the phylogenetic clustering tree under the minimum distance.
2. Construct the first-order and second-order minimum distance undirected and directed topological distance trees, and represent the topological distance using colored arcs.
3. For the SARS sequences, construct phylogenetic trees using the maximum-likelihood method first, followed by the Bayes method.

**Exercise 32.** Perform MA based on the 8–12 earliest SARS sequences. Then, analyze the network structure based on the alignment output. Compute the following results:

1. Determine the stable and unstable regions, and express these using the modulus structure.
2. Construct the phylogenetic trees using minimum distance.
3. Construct first-order and second-order minimum distance undirected and directed topological distance trees, and express the topological distance using colored arcs.
4. Based on the first-order and second-order minimum distance undirected topological distance trees, perform orthogonal mutation network decomposition, and construct the graph of the orthogonal mutation network structure.
5. Based on the graph of the orthogonal mutation network structure, and using Pagumalarvata as the source of the disease gene, explain the gene mutation process and the path of the disease infection.

**Exercise 33.** Based on the MA outputs for the ND1 gene coding region sequences of 20 mammals, construct the phylogenetic tree according to the following typical requirements:

1. Using the characteristic value of the stable regions of MA outputs, construct the phylogenetic tree using the parsimony method.
2. Construct the phylogenetic tree using the maximum-likelihood method and the Bayes method.

**Hint**

Construct the phylogenetic tree for the maximum-likelihood method and the Bayes method, using the software packages Phylip [29], Paml [111], and MrBayes [44].