Papers Summary

Qian Feng

January 25, 2018

Contents

1 Paper 1: A survey of combinatorial methods for phylogenetic networks

A survey of combinatorial methods for phylogenetic networks, Huson and Scornavacca, Genome Biol. Evol., 2011. The full pdf version please see here.

Phylogenetic trees are not suitable to describe evolutionary history when datasets involve significantly plenty of reticulate events, including horizontal gene transfer, hybridization, recombination, reassortment etc. Phylogenetic network provides an alternative. It is any graph used to represent evolutionary relationships between a set of taxa that label some of its nodes.[2]

This paper is a literature review, introducing briefly fundamental concepts about phylogenetic network and summarizing separate algorithms correspond to each type of network. Figure 1 vividly show all different types of phylogenetic network introduced in this essay.

In theory, Phylogenetic network consists of two types: unrooted phylogenetic network and rooted phylogenetic network. The former one is much more widely used than the latter one in practice, because there are many problems needed to solve in rooted phylogenetic network. First, many algorithms are not designed as a tool in real studies though they have proof-of-concept implementations; second, algorithms have impractical running times. Therefore, developing suitable methods for rooted phylogenetic network is still a unforeseeable challenge.

From another point of view, phylogenetic networks are used in two ways, the first one is as a tool for visualizing incompatible clusters/taxa, we call it "abstract", "implicit" or "data-displayed" networks; another one represents the evolutionary history including reticulate events, called "explicit" or "evolutionary " networks. In some sense, most unrooted phylogenetic networks are "abstact", however, rooted phylogenetic networks could be either abstract or explicit.

★ Pay attention to the difference among these words: Hybridization, Recombination, Mutation, Crossover, Duplication. I could not distinguish them actually. Answers given by Yao-ban:Hybridization is about the species level, Recombination is about gene level, Crossover is double recombination. Mutation is change of specific gene.Duplication could repeat, like could make the gene longer, likewise, gene loss.

2 Paper 2: Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data

Li, Na, and Matthew Stephens. "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data." Genetics 165.4 (2003): 2213-2233. The full pdf version please see here.

The patterns of Linkage Disequilibrium(LD) are the result of genetic factors and demographic history in a population[3]. Particularly, recombination plays a key role relating with the patterns of LD. For example, when a recombination occurs between two loci, it will reduce the dependence between two alleles, then reduce LD. In this article, authors propose a new statistical method to estimate the underlying recombination rate to get a better review of pattern of LD. It is also meaningful in understanding and interpreting patterns of LD and LD mapping.

Our model based on

$$P(h_1, ...h_n|\rho) = P(h_1|\rho)P(h_2|h_1;\rho), ...P(h_n|h_1, ...h_{n-1};\rho) \tag{1}$$

where $h_1, ...h_n$ denote n sampled haplotypes, ρ is the recombination parameter. In this new proposed method, we substitute an **approximation** (noted as $\widehat{\pi}$) for those conditional distributions in right term of (1), namely

$$P(h_1, ...h_n|\rho) \approx \widehat{\pi}(h_1|\rho)\widehat{\pi}(h_2|h_1;\rho), ...\widehat{\pi}(h_n|h_1, ...h_{n-1};\rho) \tag{2}$$

The right term above is what we called likelihood L_{PAC}, through maximizing this likelihood, we could get the recombination parameter ρ.

How to compute $\widehat{\pi}$ is the key point in this article. In fact, according to the appendix A, the core is the utilize of forward algorithm in HMM.

Let X_j denote which haplotype / h_{k+1} copies at site j. X_j is a Markov model on $\{1, 2, ..., k\}$ with emission probability $P(X_1 = x) = \frac{1}{k}(x \in \{1, 2, ...k\})$, the transition probability is as follows

$$P(X_{j+1} = x\prime|X_j = x) = \begin{cases} p_j + \frac{1}{k}(1 - p_j) & x\prime = x \\ \frac{1}{k}(1 - p_j) & \text{otherwise} \end{cases} \tag{3}$$

where $p_j = exp(-\frac{\rho_j d_j}{k})$, $rho_j = 4Nc_j$, N is the effective population size ,and c_j is the recombination rate per physical distance, d_j is the distance between marker j and marker $j + 1$.

Computing $\widehat{\pi}(h_{k+1}|h_1,...h_k;\rho)$ requires a sum over of all possible values of X_j, this is why we exploit forward algorithm to compute it recursively.

$$\pi(h_{k+1}|h_1,h_2,...h_k) = \sum_{x=1}^{k} \alpha_s(x) \tag{4}$$

$$\alpha_{j+1}(x) = e_{j+1}(x) \sum_{x\prime=1}^{k} \alpha_j(x\prime) P(X_{j+1} = x | X_j = x\prime) \tag{5}$$

A noticeable point in this article is that c_j has different set in different models. When the recombination is constant, ie.$c_j = \bar{c}$, we only have one parameter $\bar{c}$, in this case, we use the *golden bisection search* when maximizing the likelihood. When the recombination parameter is variable, *ad hoc* two-stage strategy is used to estimate c_j and λ_j, but we need to know the *ad hoc* two-stage approach is not guaranteed to get reliably global maximum.

★ What on earth is the *ad hoc* two-stage approach? Since authors do not pursue here, why they do not use MCMC mentioned before?

Answers given by Heejung: It is about Baysian theory. Through maximizing product of likelihood and prior distribution to get the estimation.

$$P(\mu|X_1,X_2...X_n) = \frac{P(X_1,X_2...X_n|\mu)P(\mu)}{P(X_1,X_2...X_n)} \tag{6}$$

In summary, this algorithm has several advantages and a small disadvantage. It is computationally fast, able to avoid the assumption that LD has the "block-like" structure, also able to consider all loci simultaneous rather than pairwise. However, an unwelcoming feature of this method is it does not consider the order of haplotypes which other currently available algorithms take account. Fortunately, by averaging the L_{PAC} over random orders of the haplotypes, authors find related performance is not significantly sensitive to the orders used.

3 Paper 3: Phylogenetic networks: a review of methods to display evolutionary history

Phylogenetic networks: a review of methods to display evolutionary history, David A.Morrison, Annual Research and Review in Biology, 2014

The full pdf version please see here.

Evolution involves a series of unobservable historical events, we could neither make direct observation nor perform experiment to investigate them, making phylogenetic an interesting and challenging discipline.

Sources of evolutionary novelty include vertical evolutionary processes and horizontal evolutionary processes. Phylogenetic trees are intended to solely for vertical processes, however, phylogenetic networks is more general with accommodating horizontal events. These horizontal evolutionary processes are represented by reticulation in networks. In this review paper, the author focus on the rooted phylogenetic network, even though there are a few automated methods available for constructing them. Most empirical networks are constructed either manually or by modifying the output of computer program.

Pay attention to these words:*diploid, polyploid, homoploid*

Horizontal evolutionary processes contain hybridization, introgression, HGT, recombination, viral reassortment and genome fusion. The last one is considered to be rather rare since it means the addition of whole genome from one specie to another specie.

Hybridization: Hybridization is very common in plants and a small group of animals like fish and reptiles. The new hybrid species consist same amount of genomic materials from each of the two parental species, ie 50:50 composition. Fig 2 shows the difference between homoploid and polyploid. In particular, polyploid is the only form of reticulate evolution we can construct the history by trees. From multi-labelled trees, K.T.Huber has constructed available and implementable method to construct phylogenetic network that is guaranteed to have a minimal number of interaction nodes. The core of this algorithm is merge and prune maximal inextendible subtrees and its equivalent subtrees, this process is repeated until a network is obtained that contains no repeated labeled leaves. Java package PADRE is available to visualize the network.

Introgression is not 50:50 composition because hybrid individuals back-

cross preferentially to one of the parental species.

Introgression and HGT are both the transfer of genetic materials from one specie to another, but the former one occur via sexual reproduction and latter one does not. Fig 3 also vividly show the introgression and HGT. HGT is detected by incompatibility between two or more trees for the same site of species. HGT is common among bacterias.

Reassortment means when two strains co-infect a host cell, then create a new strain by re-combing these two genetic materials.

Recombination concludes intra-genic Recombination and inter-genic Recombination. The former one represents the break-points occur within a single gene, however, the latter one can occur in different genes or non-coding space between genes. A noticeable point is crossover means double recombination. In general , genes with low level of recombination will have low levels of polymorphism, hence, recombination has important influence on genome and genetic structure in population.

Homoplasy is the development of organs or other bodily structure within different species, resemble each other and have same functions, but do not have the same ancestral origins. For instance, the wings of insects, birds and bats, are homoplastic (meaning: similar in form and structure, but not in origin).

Apparently, there is a growing need for researchers to detect and display the evolutionary networks.

Lastly, author introduced briefly current usage of different reticulate patterns. Available programs are as following: Dendroscope and SplitsTree for hybridization, SPRIT for HGT, Kwarg and SHRUB for recombination.

Feedbacks:

1. Look at paper in reference 72 since we should know more about recombination.
2. Look at more details of SHRUB software.

4 Paper 4: Reconstructing the evolutionary history of polyploid from multilabeled trees

Reconstructing the evolutionary history of polyploid from multilabeled trees. Huber, Katharina T, et al. Molecular Biology and Evolution 23.9 (2006): 1784-1791.

Polyploid species played a major role in the evolution of plants. In this paper, we focus on present all possible phylogenetic networks from a multilabeled tree that are guaranteed to have minimal number of interaction nodes.

Based on Fig.2 in this paper, we can see (b)(c)(d) are all phylogenetic networks that exhibit (a), however, we will use efficient algorithm to draw a phylogenetic network like (d) rather than (b) and (c).

From Fig.3 in this paper, we can see subtrees $T_{(u)}$,$T_{(v)}$ and $T_{(w)}$ are maximal inextendible. Fours useful concepts are subtree, equivalent, inextendible and maximal inextendible. In fact, we will focus on how to find maximal inextendible subtrees from a given MUL tree in actual algorithm. Note that the definition of *inextendible* is not clear for me, so I borrowed another one in the paper of Huber KT and Moulton to get a better understanding this terminology.

inextendible: suppose T is MUL tree, for every vertex $v \in V(T)$ that is not the root of T we denote the parent of v by $\bar{v}$, suppose T' is a sub MUL tree with vertex v, we say T' is *inextendible* if there existed another sub MUL tree T'' with root vertex w so that T'' is isomorphic to T', and $T(\bar{v})$ is not isomorphic to $T(\bar{w})$.

So there would be a **contradiction** in the statement of inextendible definition. Take a look at Fig.3 again, whether on earth each subtree having leaves labeled with "b" and "c" is inextendible or not ?

The core of this algorithm showed in Fig.4 is to merge and prune maximal inextendible subtrees and its equivalent subtrees, this process is repeated until a network is obtained that contains no repeated labeled leaves.Unfortunately, I could understand how do they find the maximal inextendible subtrees by height list H and code $c(v)$ in the initial step.

A noticeable limitation is when MUL tree contains polytomies showed in Fig.6,using this presented constructing methods would lead to several different phylogenetic networks.

5 Paper 5: Joint Bayesian Estimation of Alignment and Phylogeny

Benjamin D. Redelings, Marc A. Suchard; Joint Bayesian Estimation of Alignment and Phylogeny, Systematic Biology, Volume 54, Issue 3, 1 June 2005, Pages 401-418.
The full pdf version please see here.

In this paper, authors propose a novel model to estimate multiple sequence alignment , phylogenetic tree reconstruction and their support simultaneously. They operate in Bayesian framework and use MCMC methods.

In general, researchers first need to finish multiple sequence alignment, and then use it to reconstruct the phylogeny. However, if the alignment contains ambiguous regions particular like in distantly related sequences, this would lead to inaccurate result. A normal technique is to remove these regions, hence, leading to a loss of a large fraction of informative sites. Due to this, researchers come up with a simple method: split ambiguous columns into groups of residues in which homology is unambiguous, then, place them in separate columns, yet, it is still worth noting that identifying these ambiguous regions is too subjective.

There are a number of techniques are developed to get a better use of ambiguous alignment. One technique, known as elision, concatenates a set of near-optimal alignments into a larger alignment and use them for reconstructing phylogeny. However, elision treat all the near optimal alignment equally instead of weighted, as a consequence, it may be not so good. Another technique, known as optimization alignment, involves estimation of alignment and phylogenies simultaneously with parsimony framework. One problem of optimization alignment is the measure of uncertainty is hard to obtain since standard bootstrap couldn't be applied due to the dependent alignment columns. Method in this paper not only weighted the alignment naturally, but also assess the confidence using posterior probability.

Advantages of joint estimation are as follows: one thing is joint estimation doesn't require extra guide tree, this contrast with alignment with progressive alignment are biased and need a guide tree. Another thing is joint estimation could provide more accurate substitution and indel (inseration/deletion) models, in scoring alignment by extended substitution model, in using shared indels to group taxa on a tree.

In developing indel model, there is a simplified assumption: indel event occurs independent on each branch. The reason of this is it allows us to

use a simple pair-HMM to model the alignment. In addition, insertions and deletions are equally likely and sequence lengths do not grow or shrink over time. In substitution model, evolution is independent across columns of f and each branch.

MCMC needs to sample from posterior distribution of the alignment, phylogeny and model parameters given only unaligned sequences. In this paper, researchers introduce a new transition kernel to resample the topology and alignment that improves mixing efficiency, allowing chains to converge even when start with an arbitrary alignment.

Multiple sequence alignment A is actually a matrix f, it specifies which letters from sequences are homologous by arranging homologous letters into the same column in this matrix.

Methods
This is an annotation list:

Item	Name	Meaning
1	Y	a set of n homologous molecular sequences
2	A	multiple sequence alignment
3	τ	unrooted tree topology
4	T	branch length
5	Θ	substitution process parameters
6	Λ	indel process parameters
7	N	total number of nodes in τ:$N = 2n - 2$
8	B	total number of branches in τ:$B = 2n - 3$
9	γ	distribution of ancestral letters at the root node
10	b	every branch
11	$\rho(b)$	parent branch for b
12	$n(b)$	node in τ shared by the branch b and $\rho(b)$

Therefore, the whole state space Ω is composed of points: $\omega = (Y, A, \tau, T, \Theta, \Lambda)$.

5.1 Substitution model

Likelihood $P(Y|A, \tau, T, \Theta, \Lambda)$ is given in substitution model, before getting this result, we need to accomplish two tasks, first one is to specify how A arranges Y into matrix f, next one is to specify the probabilistic model on the columns of f. Figure 1 in the original paper solves above first task, the tuples at the leaf node within a column in matrix f are from a multinomial distribution which addresses second task mentioned before. An worth noting thing is there is *Felsenstein wildcards* in matrix f which represents

internal nodes that are present but unobserved denoted by $*$. The full likelihood is by multiplying likelihood of each column using peeling algorithm, meanwhile Felsenstein wildcards and gaps are treated as missing data.

A Markov model is reversible equivalent to hold the detail balance equation:

$$\pi_i p_{ij} = \pi_j p_{ji} \tag{7}$$

5.2 Gap model

We describe prior of indel process parameters Λ in gap model. In gap model, A could be represented as a tuple pairwise alignments: $(A^{(1)}, A^{(2)}, ..., A^{(B)})$, we replace the alignment prior (equation(11)) with standard prior (equation(10)), then after calculating this modified posterior, Gibbs sample from it using DP programs, detailed DP algorithm is provided in Appendix.

Three parameters are used in pair-HMM, including δ,$\epsilon$and ζ,parameter δ refers to the probability of an indel in either sequence, we assume double-exponential distribution on the approximate log odds of δ; ϵ refers to the probability of extending an existing gap, and exponential distribution is assumed, for ζ, it means the transition probability from any state to the end state, in our example, $\zeta = 1/1000$.

5.3 MCMC sampling process

In MCMC sampling process, researchers employ a random scan Metropolis-within-Gibbs approach. In every iteration, they attempt to sample from every model parameters at least once. Indel parameters are resampled more frequently than substitution parameters. NNI is used to update topology τ, it will move across every internal node at least once per iteration.

In whole, topology τ is updated by a number of MH steps, each alters only part of the topology. Once a topology is chosen, the internal nodes are resampled from the DP matrix, after this step, alignment A and topology τ is resampled again. The MH acceptance probability is given in equation (12). This is an 1D DP problem.

Regarding to alignment sampling, traditional two MCMC transition kernels are not efficient enough in some circumstances shown in Figure 4, researchers decide to resample the alignment along a branch and the sequence at one end of this branch in the same step, this is a 2D DP problem.

Alignment uncertainty plot (AU) is introduced to depict the alignment variability, it is drawn from the posterior alignment, and provides a valuable tool to assess alignment ambiguity.

We show this modified MCMC algorithm converges to equilibrium distribution more quickly than previously available MCMC transition kernel. This allows us to choose randomly one alignment rather than start from an estimated alignment via other software like ClustalW, which is the first advantage, another benefit is that it would decrease burn-in time and less autocorrelation. In order to assess the convergence of continuous parameters, Gelman-Rubin R statistics is employed, results suggesting each chain converges to the same distribution, 95% Baysian credible intervals about posterior probability (PP) are also given in table 2.

5.4 BAli-Phy show

BAli-Phy is a C++ software providing samples from posterior of alignment and phylogeny model. I will show it step by step to interested readers. This software please click here.

Firstly install the BAli-Phy using homebrew based on its User's Guide in this software website. We will run commands in terminal. Type this command to check its version: bali-phy –version

Next step is to install programs used for viewing the results
1. Tracer : MCMC parameter, diagnostic viewer.
2. FigTree : Phylogeny Viewer
3. SeaView : Alignment viewer.

Then a quick start to run BAli-Phy is to type in two following commands:
bali-phy /examples/sequences/5S-rRNA/5d.fasta –iter=150
bp-analyze 5d-1/
The output results are in a separate file named "5d-1" which is in root directory (/fengqian), including

Item	Name	Meaning
1	C1.log	Numeric parameters: indel and substitution rates, etc. Opened by Tracer
2	C1.trees	Tree samples: one sample per line, in Newick format. Opened by FigTree
3	C1.Pp.fastas	Sampled alignments for partition p including ancestral sequences
4	C1.out	Iteration numbers, probabilities, success probabilities for transition kernels

and other little files.

Last, to summarize the output, type this to find majority consensus tree, note the dir has changed to 5d-1.
trees-consensus C1.trees > c50.PP.tree

To compute the maximum a posteriori tree, input :
trees-consensus –skip=10% C1.trees –map-tree=MAP.tree

Checking topology convergence use:
trees-bootstrap C1.trees
it will show us the PP and LOD values regarding to each topology as well.

Anyway, even though this framework enjoys great advantages in estimating alignment and phylogeny at the same time, there exist some limitations as well. First is the parameters prior are not sufficient for biological meaning. Another shortcoming is indel process parameters are the same along each branch, which contrasts with ClustalW. In ClustalW, there are a function of branch lengths.

It is a challenge for me to understand equation (4) and (6) in this paper fully.

PS: I am learning MCMC from Chib Siddhartha and Edward Greenberg (1995)'s paper, a powerful algorithm, it seems pretty interesting. ★

6 RDP4 note

RDP4 is the latest version of recombination detection program in a set of aligned sequences. RDP4 could not do multiple sequence alignment, but from its instruction manual, it provides three reliable sequence alignment tools: ClustalX/W, MUSCLE, POA for small, medium and large datasets respectively(small means fewer than 100 sequences, large means more than 100 sequences).

6.1 Compile a good dataset

Firstly, we need to compile a good dataset. Although there is no formula to tell us the optimal numbers and lengths of sequences for optimal recombination detection, some procedures are needed to ensure a reasonable good dataset.

6.2 Make a good alignment

After getting a suitable datasets, next step is to make a good alignment, which is essential for recombination analysis. It is not recommended that any pair of sequences share less than 60% nucleotide sequence identity, ideally this value is greater than 70%. Multiple sequence alignment tools will occasionally make some alignment errors, that's why we need to realign subsections of alignment to rectify these errors.

In general, after making a preliminary alignment of the sequences, if these are small sequences, we could use an alignment editor such as MEGA or IMPALE to check the accuracy of completed alignment by eyes, if they are large datasets, using the sub-sequence realignment tool in MEGA or IMPALE with different alignment parameter settings. It is strongly recommended that any unalignable (or just barely alignable) tracts be either deleted from the alignment or shifted/staggered.

6.3 Prelimnary scan

We can click "open" and load alignment files, RDP4 could recognize these different file formats: FASTA, PHYLIP, GDE, CLUSTAL, GCG,NEXUS,MEGA, DNAMAN, .pdb.

We need to set up some parameters before general scanning. First one is to specify whether the sequence being examined is linear or circular. The following picture is an example of circular sequence.

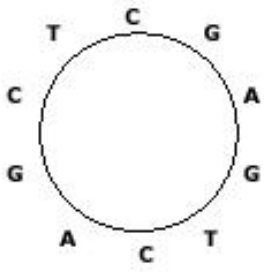

Figure 1: This graphic is originally from here

We always read a circular sequence in the clockwise direction, then it should be "CGAGTCAGCT" in the above example. Of course there can be many linear sequences that are obtained from such a circular sequence, by cutting any place of the circular sequence.

Move to "Analysis Sequences Using" in "general " button, it is strongly recommended to use the default setting. Pay attention to the Botscan and Siscann, there are two boxes in front of it separately. By default these two will automatically check the recombination signals detected by other methods,if we click the first box, then it will force their use to explore new signals, but be warned that it will increase analysis time dramatically. By the way, LARD method is also only used for checking signals and suitable for less than pretty small dataset (less than 20 sequences).

After general settings, click "Run" button in the command button panel.

6.4 Refine Preliminary recombination

It is important to be aware that RDP4 can get things horribly wrong, such as inaccurate identification of breakpoint positions. Unfortunately, there is no automated tools to judge whether our results are true.

In order to check the accuracy of estimation breakpoints estimation, best graphs are shown in the bottom left panel by choosing "Check using " "MAXCHI"and "CHIMERA". Breakpoints are displayed in the peak of these two curves. However, if the peaks doesn't match the border of recombinant, this doesn't mean this inferred positions are wrong, it does mean there is a degree of uncertainty regarding this position. By the way, considering "Confirmation Table" in Recombination Info at top right panel is also a good way to assess the accuracy of estimation.

Remember this program could crash at any time so we should regularly save our results.

7 Book 1: Phylogenetic networks: concepts, algorithms and applications.

Huson, Daniel H., Regula Rupp, and Celine Scornavacca. Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press, 2010. The full pdf version please see here.

Reassortment is the mixing of the genetic material of a species into new combinations in different individuals. It is particularly used when two similar viruses that are infecting the same cell exchange genetic material. [1]

There are several interesting concepts in graph theory. Please look at the Figure 1.1, 1.2 and 1.3.

1. The *degree* of node u is the sum of its indegree and outdegree.
2. A *biconnected components* is the maximal subgraph that is induced by set of edges and doesn't contain a cut node. A good example is in Figure 1.3.
3. A graph $G = (V, E)$ is called *bipartite* if and only if its set of nodes can be partitioned into subsets V_1, V_2,with $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \emptyset$, such that for every edge $e \in E$,one of the endpoints lies in V_1, another endpoint lies in V_2.
   Explain Exercise 1.2.3.
4. Two nodes v and w are *imcomparable*, if neither node is lower than other; similarily, two edges e and f are *imcomparable* if neither is lower than other;
5. Different *traversals* give rise to different orders in which nodes are examined. Pay attention to preorder, postorder and breadth-first traversal in Figure 1.7. In particular, breadth-first traversal, please reference here.

PS: Nodes are also called vertices, edges are also called branches or arcs. Bifurcating Tree = Resolve Tree = Binary Tree

Let $\chi = \{x_1, x_2, ...x_n\}$ be a set of taxa, a *cluster* is any subset of χ,excluding the empty set $\emptyset$ and full set χ. The ultimate goal of phylogenetic analysis is to compute a set of clusters on χ such that each cluster is monophyletic(also called clade. Monophyletic group contains all descendants of the common ancestor and the ancestor itself).

A *split* is any bipartitioning of χ into two non-empty subsets A and B of χ, such that $\chi = A \cup B$ and $A \cap B = \emptyset$.

In phylogenetic analysis, a set of taxa $\chi = \{x_1, x_2, ...x_n\}$ is often represented by a set of molecular sequences $A = \{a_1, a_2, ...a_n\}$ where a_i comes from taxon x_i and correspond to some specific genes or locus. We also need to ensure that the sequences are **homologous**, that is, have evolved from a common ancestor sequence.

In *Pairwise sequence alignment* , with the help of substitution matrix, for example the BLOSSUM matrix, which assigns empirically score, we could calculate the score of each pair of residues and then sum over scores among all pairs would be the score of whole alignment.

Sequence are often aligned by inserting gaps into each sequence shown in Figure 2.6 such that all sequences have same length m, forming a *multiple sequence alignment* of length m. Our goal is to find a multiple sequence alignment that achieves the optimal score according to an appropriate score scheme. **Progressive method** as a heuristic approach, is used to align multiple sequences, its outline is shown in Figure 2.7. The core is to align a pair of similar sequences into *profiles*, then align profiles into final multiple sequence alignment.

Let M be a multiple sequence alignment on χ, each column of M is called a character, each symbol that occurs in this column is called a character state.

Now we are introducing some basic concepts and main methods for inferring phylogenetic trees.

Phylogenetic trees are usually computed from molecular sequences. They not only could uncover the relationship between different species or taxa, but also have many other applications. For instance, they are used to determine the age and the rate of diversification. In sequence-analysis method, they are allowed *phylogenetic footprinting*.

In practice, there are two types of analysis after the initial multiple sequence alignment: distance-based analysis and sequence-based one. Its outline is shown in Figure 3.1 at page 24.

Definition Phylogenetic Tree
Given a set of taxa χ, this is a phylogenetic tree $T = (V, E)$, its all nodes have degree $\neq 2$, together with a taxon labeling $\lambda : \chi \to V$that assigns actually one taxon to every leave and none to internal nodes.

From a theoretical and algorithmic point of view, unrooted phylogenetic trees are much more easier than rooted ones, however, in biology, rooted phylogenetic trees are usually more of interest. A phylogenetic tree is called an edge-weighted tree if we are given a map ω that assigns a non-negative weight or length $\omega(e)$ to every edge e of the tree. In drawings, we usually use length of the edge to indicate the scale rather than write the lengths explicitly next to edges.

Jukes-Cantor model tells us the probability formula of change during time t or along the edge, given the mutation rate. This model of DNA evolution assumes the fours bases (A,C,G and T) occur with equal frequencies(0.25) and change from one base to another occurs at the same rate. If we relax the conditions, for example, let the bases occur at different and arbitrary rates (although they have to sum to 1), change rates in transitions and transversions, then we could get more generate model, anyway, they are both special cases of general time reversible model.

Classical phylogenetic trees construction approaches consist of two following types:

* **Sequence-based method** usually searches for best phylogenetic tree which can optimally explain the given multiple sequence alignment M. We discuss the three main approaches about it: maximum parsimony, ML and Bayesian inference.

* **Distance-based method** usually constructs phylogenetic tree from a given a distance matrix D.

7.1 Maximum parsimony

Maximum parsimony method is to look for a phylogenetic tree that explains the given set of aligned sequences using a minimum number of evolutionary events.

The *parsimony score* of T(tree) and M(given multiple sequence alignment) is defined as:

$$PS(T, M) = \min_{\alpha} \sum_{\{x,y\}} diff(x, y) \tag{8}$$

where diff() function is known as *hamming distance* between sequence $x = (x_1, x_2, ...x_m)$and sequence $y = (y_1, y_2, ...y_m)$ that describes the difference of $x$and$y$

$$diff(x, y) = |\{i|x_i \neq y_i\}| \tag{9}$$

The minimum is taken over all possible assignments α that make the sequences of length m to be the internal nodes, summation is taken over all possible pairs of x and y that are assigned at opposite end of edge of T .

This task of computing parsimony score is the known *small parsimony* problem. For bifurcating tree, *Fitch algorithm* is used to calculate efficiently, in more general settings, *Sankoff's algorithm* can be applied. Let's describe Fitch algorithm. Assume we are given a multiple sequence alignment M and a bifurcating tree T on χ, we need to score each character (that is, column of the alignment) separately, and then obtain the parsimony score $PS(T, M$ by summing over all characters.

References

[1] Daniel H Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, 2010.

[2] Daniel H Huson and Celine Scornavacca. "A survey of combinatorial methods for phylogenetic networks". In: *Genome biology and evolution* 3 (2011), pp. 23–35.

[3] Na Li and Matthew Stephens. "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". In: *Genetics* 165.4 (2003), pp. 2213–2233.