Papers Summary

Qian Feng

December 19, 2017

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 backcross 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, David A.Morrison, Annual Research and Review in Biology, 2014

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 $T\prime$ is a sub MUL tree with vertex v, we say $T\prime$ is *inextendible* if there existed another sub MUL tree $T\prime\prime$ with root vertex w so that $T\prime\prime$ is isomorphic to $T\prime$, 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 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 $\lambda : \chi \rightarrow V$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.

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.

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.