MODELING SPECIES-GENES DATA FOR EFFICIENT PHYLOGENETIC INFERENCE

Wenyuan Li and Ying Liu*
Department of Computer Science, University of Texas at Dallas
Richardson, TX 75083, U.S.A.
**Email: ying.liu@utdallas.edu*

In recent years, biclique methods have been proposed to construct phylogenetic trees. One of the key steps of these methods is to find complete sub-matrices (without missing entries) from a species-genes data matrix. To enumerate all complete sub-matrices, [17] described an exact algorithm, whose running time is exponential. Furthermore, it generates a large number of complete sub-matrices, many of which may not be used for tree reconstruction. Further investigating and understanding the characteristics of species-genes data may be helpful for discovering complete sub-matrices. Therefore, in this paper, we focus on quantitatively studying and understanding the characteristics of species-genes data, which can be used to guide new algorithm design for efficient phylogenetic inference. In this paper, a mathematical model is constructed to simulate the real species-genes data. The results indicate that sequence-availability probability distributions follow power law, which leads to the skewness and sparseness of the real species-genes data. Moreover, a special structure, called "*ladder structure*", is discovered in the real species-genes data. This *ladder structure* is used to identify complete sub-matrices, and more importantly, to reveal overlapping relationships among complete sub-matrices. To discover the distinct ladder structure in real species-genes data, we propose an efficient evolutionary dynamical system, called "*generalized replicator dynamics*". Two species-genes data sets from green plants are used to illustrate the effectiveness of our model. Empirical study has shown that our model is effective and efficient in understanding species-genes data for phylogenetic inference.

1. INTRODUCTION

Phylogenetic inference can be defined as the process of determining estimated evolutionary history by analysis of a given data set [18]. The evolutionary history of genes and species can be described by a phylogenetic tree [12, 7]. It is widely accepted that amino acid and/or DNA sequences produce a tree closest to the true tree [6, 15, 8]. As the amount of molecular sequence data available rapidly increases, it has spurred a number of phylogenetic analysis across the tree of life [16]. In general, the data prepared for phylogenetic analysis is in the form of species-genes matrix, where genes refer to any set of homologous sequences, whether protein coding or not [17, 5]. As species-genes matrix indicates whether there exist sequences for any species and gene, it is also called sequence availability matrix. Ideally, this matrix is complete, which means that every species has been sequenced for every gene in the matrix. However, as pointed out by [17], a few species have been sequenced for many genes; a few genes have been sequenced for many species; but most of the potential data available for phylogenetic purposes is still missing. Therefore, species-genes matrices derived from the available sequence data are "*sparse*" and "*uneven*" [14, 11, 2, 17]. The sparseness and skewness of species-genes data have posed serious challenges for the available phylogenetic methods and strategies of constructing trees [16, 19].

Recent studies have shown that concatenating multiple sequences from the same species can improve the accuracy of phylogenetic inference [11, 2, 17, 5]. Given a large species-genes matrix, Sanderson *et al.* (2003) developed an exact algorithm to find all complete sub-matrices (without missing data). Once all the complete sub-matrices are discovered, an effective strategy for constructing phylogenetic tree is to concatenate the sequences of all the genes in the complete sub-matrix. The whole process is illustrated in Fig. 1. In this process, an important step is the discovery of all complete sub-matrices. In graph theory, the species-genes matrix $W = (w_{ij})_{m \times n}$ (as right part of Fig. 1) can be represented as a bipartite graph $\mathcal{G}(S, G, W)$, in which there are two vertex sets $S = \{s_1, \ldots, s_m\}$ (s_i rep-

*Corresponding author.

resents the i-th species) and $G = \{g_1, \ldots, g_n\}$ (g_j denotes the j-th gene), and the edge between s_i and g_j exists if $w_{ij} = 1$ (the gene g_j is sequenced for the species s_i) and otherwise if $w_{ij} = 0$. Therefore, a complete sub-matrix of W corresponds to a complete subgraph of $\mathcal{G}$, typically called 'biclique' in graph theory [a]. Therefore, in a biclique, all genes are sequenced for all species. In essence, the discovery of all complete sub-matrices is equivalent to an NP-complete graph problem, known as "biclique enumeration" [1, 17, 16]. The running time of the "biclique enumeration" algorithm proposed by Sanderson *et al.* (2003) is exponential and may take a long time to analyze large data sets. Furthermore, it generates a large number of bicliques, many of which may not be used for tree reconstruction. Hence, it is time-consuming for phylogeneticists to determine which bicliques can build meaningful phylogenetic trees.

Although the sparseness and skewness of species-genes data is a curse for biclique enumeration, they may become a blessing for phylogenetic inference if we study and take advantage of them. Therefore, in this paper, we focus on quantitatively studying and understanding the characteristics of real species-genes data, which can be used to guide new algorithm design for efficient phylogenetic analysis. We firstly construct a mathematical model to simulate real species-genes data. Then some underlying and special features or structures, such as "ladder structure", can be discovered in real species-genes data through the model. This ladder structure can be used to identify complete sub-matrices, and more importantly, to reveal the distinct overlapping relationships among complete sub-matrices. Finally, to discover the ladder structure in real data, we propose an efficient evolutionary dynamical system, called "*generalized replicator dynamics*".

The rest of this paper is organized as follows: we firstly propose a model in Section 2 to study the species-genes data. Based on this model, characteristics of the real-world species-genes data can be quantitatively investigated. Two conclusions are drawn when we use this model to analyze two real species-genes data sets collected from green plants. In Section 3, we formulate the discovery of ladder structure as a maximization problem. To approach this problem, in Section 4, we generalize a well-known population dynamics in the evolutionary biology, replicator dynamics, to the general matrix, i.e. species-genes data matrix, for efficiently estimating our model [9, 10]. We call this new dynamics "*generalized replicator dynamics*". Empirical results in Section 5 show our model can effectively and efficiently build phylogenetic trees by estimating its distributions. Finally, conclusions and future works are presented in Section 6.

2. MODEL OF SPECIES-GENES DATA

Before introducing the model of species-genes data, we firstly review two characteristics recently observed by phylogeneticists [17, 16]:

(1) **Sparse and uneven sequence availability distribution:** as shown in Fig. 2(d) that is an excerpt from [16] and Fig. 2(e), these two matrices are very sparse and uneven. "*Many sequences are available for a few species and a few heavily sampled genes are available for many species*". Moreover, these two figures show "*the most heavily sampled corner of the species-genes matrix and the remainder of the matrix is even more sparse*" [5].

(2) **Many overlaps of bicliques:** as observed and reported in [17], "*many of bicliques overlap, and for any given biclique there are generally bicliques which have either slightly more species and slightly fewer genes, or slightly more genes and fewer species*".

The first characteristic is an empirical observation of a global data distribution in the whole species-genes data and the second one indicates relationships among bicliques. However, they are only qualitative and rough view of the species-genes data and thus may not provide further useful guidance and insights for data analysis algorithm design. Therefore, we build a model to quantitatively analyze species-genes data and advance our understanding of species-genes data from qualitative observations to quantitative investigations.

[a] In the rest of the paper, for simplicity, we use the term 'biclique' to denote 'complete sub-matrix'.

As a few species are sequenced for many genes and a few genes are sequenced for many species in the real-world species-genes data, in the model, we assign each species $s_i \in S$ a Sequence-Availability (shortly denoted as SA) probability value p_i^S and each gene $g_j \in G$ a SA probability value p_j^G as well. The greater the sequence-availability probability of s_i or g_j is, the larger the number of genes or species the corresponding species s_i or gene g_j is sequenced for. Therefore, two sets of SA probabilities are used in the model: one is species SA probability distribution $\mathbf{p}^S = \left(p_1^S, p_2^S, \ldots, p_m^S\right)$, and the other is genes SA probability distribution $\mathbf{p}^G = \left(p_1^G, p_2^G, \ldots, p_n^G\right)$, where m is the number of species ($m = |S|$), n is the number of genes ($n = |G|$), and $p_i^S, p_j^G \in [0, 1]$. Then the species-genes data matrix $W = (w_{ij})_{m \times n}$ (or sequence availability matrix) can be simulated by the SA probability distributions $\mathbf{p}^S$ and $\mathbf{p}^G$ of m species and n genes in the model described as follows,

Model of species-genes data (or sequence availability data)

Input: *species SA distribution* $\mathbf{p}^S$ *and genes SA distribution* $\mathbf{p}^G$*;*

Output: *species-genes data* $W_{m \times n}$*.*

1. *generate a uniformly random value* $p \in [0, 1]$*;*
2. *the sequence is available for the species* s_i *and the gene* g_j *(i.e.,* w_{ij} *is set 1), if* $p \leqslant p_i^S$ *or* $p \leqslant p_j^G$*; otherwise, the sequence is missing for* s_i *and* g_j *(i.e.,* w_{ij} *is set 0).*
3. *iterate step 1 and 2 for all species and genes.*

In this model, the larger the p_i^S and p_j^G are, the more possible the species s_i is sequenced for the gene g_j (i.e., the corresponding w_{ij} is 1). Therefore, the SA probability of a species (or gene) determines if this species (or gene) is sequenced more or less.

With this model, we are interested in quantitatively answering two questions that are already partially and qualitatively answered in the observed characteristics mentioned above by Sanderson *et al.* (2003): (i). "*what are the sequence availability distributions of species and genes,* $\mathbf{p}^S$ *and* $\mathbf{p}^G$*?*". We already know they are skewed from the first characteristic aforementioned and further want to know "*how skew are they?*" (ii). "*what are the relationships of bicliques?*" We already know many bicliques overlap in the way described in the second characteristics aforementioned. But we further want to know "*structures of these overlapping bicliques and to what extent they overlap?*"

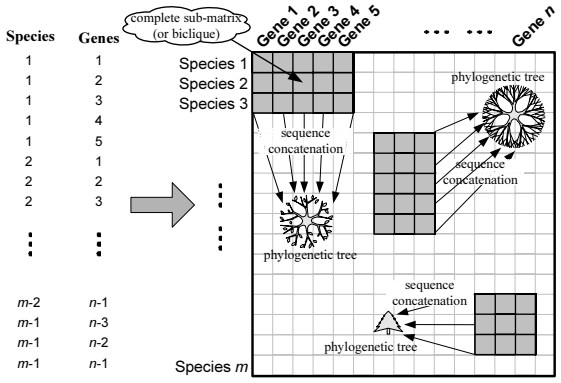


Fig. 1. Construction process of phylogenetic trees from species-genes data: from a list of genes and the species for which sequences of those genes are available (as left), to species-genes matrix (as right), complete sub-matrix(or biclique) discovery, sequence concatenation in biclique, and phylogenetic trees construction.

Therefore, to answer the first question, a useful method is to try different statistical distributions as SA distributions of species and genes, i.e., $\mathbf{p}^S$ and $\mathbf{p}^G$ in the model, and to compare these simulated species-genes data with the real data in terms of both matrix similarity and sequence numbers' distributions [b]. After answering the first question by determining which statistical distribution the real species-genes data follows, we can simulate species-genes data using the model, and then more insights and understanding of species-genes data for the second question may be obtained from the study of the simulated species-genes data. In the following, we report our results regarding to the above two questions one by one.

Conclusion 1: sequence availability probability distributions of species and genes follow power law. To answer the first question and determine what SA probability distributions in real data are, we utilized three types of statistical distributions, (a) uniform, (b) normal and (c) power

[b]The sequence number of a species (or gene) is the total number of sequences available for this species (or gene) across all genes (or species). Specifically, the sequence number $d(s_i)$ of the i-th species s_i is $d(s_i) = \sum_{j=1}^{n} w_{ij}$. Similarly, the sequence number $d(g_j)$ of the j-th gene g_j is $d(g_j) = \sum_{i=1}^{m} w_{ij}$.

law [c]. Their skewness increases from type (a) to type (c). In this paper, we simply let $\mathbf{p}^S = \mathbf{p}^G$ for only emphasizing the difference between distribution types. In practice, to make the simulated data closer to the real data, different parameters of the distribution can be tested for $\mathbf{p}^S$ and $\mathbf{p}^G$. The corresponding simulated species-genes matrices of the distributions (a), (b) and (c) are shown in Fig. 2(a), 2(b) and 2(c), respectively. For comparison, two real-world species-genes data matrices are also drawn in Fig. 2(d) and 2(e). All matrices in Fig. 2 are rearranged by the decreasing order of sequence numbers of species and genes. It can be clearly observed that, with the increase of distribution's skewness from type (a) to type (c), the first characteristic aforementioned becomes more and more obvious, e.g., matrices are more and more sparse and skewed. In addition to comparing simulated and real data by matrix similarity, we can also compare their sequence numbers' distributions of species and genes. They are plotted in the form of log-log cumulative distribution [d]. We found that in two real data sets as shown in Fig. 2(d) and 2(e), species and genes' sequence numbers' distributions in log-log form are roughly straight lines and thus follow power law. When observing three simulated data from type (a) to type (c), their sequence numbers' distributions evolve from curves to lines. Therefore, the data simulated by the model with the statistical distribution type (c) is much closer to real species-genes data. Hence we draw the conclusion that sequence availability probability distributions of species and genes in real data follow the statistical distribution type (c) – power law.

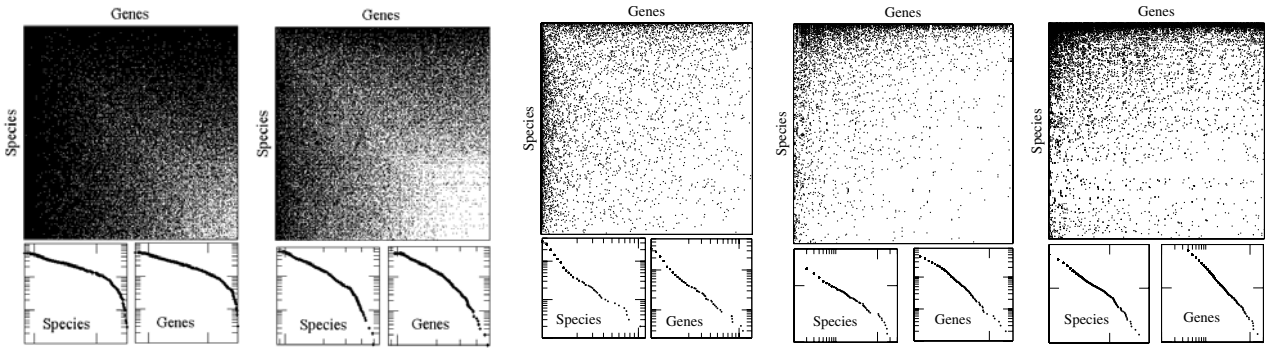
Although this is not a surprising result, it gives us an idea of how skewed real species-genes data is. Furthermore, it provides us a way to study the structures and properties of real species-genes data. Next, we employ this model with the power law SA distributions of species and genes to study the second question.

Conclusion 2: the most distinct overlapping structure of bicliques is a ladder structure. Based on Conclusion 1, we used our model with the power law SA distributions of species and genes to generate a species-genes data matrix with 20 species and 20 genes. After rearranging the simulation matrix with the decreasing order of $\mathbf{p}^S$ and $\mathbf{p}^G$ as shown in Fig. 3, a distinct structure is revealed to the left-top corner of W. Bicliques are easily identified and the overlapping relationships among bicliques are also clearly shown in Fig. 3. Each box framed by dotted lines is a biclique in Fig. 3. Therefore, this distinct structure is useful for not only locating bicliques, but also intuitively revealing the overlapping relationships among these bicliques. In this structure, bicliques overlap in the way that confirms what Sanderson *et al.* (2003) described, "*many of bicliques overlap, and for any given biclique there are generally bicliques which have either slightly more species and slightly fewer genes, or slightly more genes and fewer species*". As this structure is like a ladder, we call it "*ladder structure*". It can also be found that the sequence availability probability actually plays the role of measuring the contribution of each species and gene to the ladder structure. Therefore, although some species or genes have small sequence numbers, their sequence availability probabilities are very high. For example, the 10th species has only 4 sequences available, much smaller than the last (20th) species with 8 sequences. Hence, the ladder structure can be identified in the left-top corner of the species-genes matrix rearranged by the decreasing order of the estimated $\mathbf{p}^{S*}$ and $\mathbf{p}^{G*}$, not by the decreasing order of sequence numbers' distributions. As this small-size simulated species-genes data in Fig. 3 is generated by the same model with the same SA distributions of species and genes and parameters as the large-size simulated data in Fig. 2(c), it is reasonable to infer that, in real data, "***the most distinct overlapping structure of bicliques is ladder structure***" and "***the distinct ladder structure in real data can be discovered by the estimated SA distributions*** $\mathbf{p}^{S*}$ ***and*** $\mathbf{p}^{G*}$".

From the above model analysis of species-genes data, two conclusions are useful for effective and efficient phylogenetic tree inference. Especially, the lad-

[c]Power law distribution follows the rule $P(x) = \alpha x^{-\beta}$. It can be seen as a straight line on log-log figure. More detail refers to `http://en.wikipedia.org/wiki/Power_law`.

[d]For a species (or gene) sequence number k, how many species (or genes) have sequence number higher than k.



(a) Uniform (b) Normal ($\mu = 0, \sigma = 1$) (c) Power law ($\alpha = 1, \beta = 1$) (d) GenBank data (e) SwissProt data

Fig. 2. Comparison of matrices and sequence numbers' distributions among real and simulated species-genes data. In species-genes matrices, a dot indicates the existence of a sequence for that species and genes. And species are sorted vertically by their number of sequences, and genes are sorted horizontally by the number of taxa for which they have been sequenced. Sequence numbers' log-log cumulative distributions of species and genes are plotted in the bottom of each species-genes data matrix. In the plot, the x-axis is the sequence number of a species (or gene) k, and the y-axis is the number of how many species (or genes) have sequence number higher than k. **(a), (b), and (c):** simulated data by the model with uniform, normal and power law distributions of sequence-availability probabilities of species and genes. **(d) GenBank data:** The most-represented species (*Arabidopsis thaliana*) is at the top, and the most heavily sequenced gene (*rbcL*) is on the left. **(d) SwissProt data**. Images in (d) and (e) show the most heavily skewed corner of the species-genes matrix and the remainder of the matrix is very sparse.

der structure in the second conclusion, which was not discovered before, can be helpful for biclique discovery and tree reconstruction. Once the ladder structure is discovered, bicliques can be easily identified (as illustrated in Fig. 3). In the rest of the paper, we will focus on discovering the distinct ladder structure in real species-genes data by estimating its SA distributions.

3. PROBLEM OF ESTIMATING SA PROBABILITY DISTRIBUTIONS

According to the model introduced above, the estimated sequence-availability distributions of species and genes are key to the discovery of the ladder structure. To estimate SA distributions in the real species-genes data, we introduce an availability probability $p_{ij} = p_i^S p_j^G$ for the sequence existence in the i-th species s_i and the j-th gene g_j. Then, a sequence availability probability matrix $P = (p_{ij})_{m \times n}$ can be constructed. To measure how well this sequence availability probability matrix P approximates the actual sequence availability matrix (real data W), we introduce a function called "*Accumulated Probability Function of Sequence Availability*", denoted as $P(\mathbf{p}^S, \mathbf{p}^G, W) = \sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} p_i^S p_j^G$, to count all the availability probabilities of sequences existing in the real species-genes data W. Intuitively, $\mathbf{p}^{S*}$ and $\mathbf{p}^{G*}$ that can maximize $P(\mathbf{p}^S, \mathbf{p}^G, W)$ will make the matrix P approximate the matrix W to the maximal extent. Therefore, by maximizing the function $P(\mathbf{p}^S, \mathbf{p}^G, W)$, we can obtain the estimated $\mathbf{p}^{S*}$ and $\mathbf{p}^{G*}$ for the real species-genes data W. Then the distinct ladder structure hidden in the real species-genes data W can be discovered in the left-top corner of W reordered by the decreasing orders of $\mathbf{p}^{S*}$ and $\mathbf{p}^{G*}$. In practice, to limit the value range of the function $P(\mathbf{p}^S, \mathbf{p}^G, W)$, the constraints on $\mathbf{p}^S$ and $\mathbf{p}^G$ are added to the problem formulation. The constraints are only the normalization of $\mathbf{p}^S$ and $\mathbf{p}^G$ and thus will not affect the discovery of ladder structure. It is formally expressed as follows,

$$\arg \max_{\mathbf{p}^S \in \Delta_+^m, \mathbf{p}^G \in \Delta_+^n} P(\mathbf{x}, \mathbf{y}, W) \tag{1}$$

where $\Delta_+^n = \left\{ \mathbf{x} \in \mathbb{R}^n \middle| \sum_{i=1}^{n} x_i = 1, \text{and } x_i \geqslant 0 \ (i = 1, \ldots, n) \right\}$ denotes a superplane in n-dimensional non-negative vector space.

In the next section, we will propose an efficient algorithm to efficiently maximize $P(\mathbf{p}^S, \mathbf{p}^G.W)$.

4. ALGORITHM OF ESTIMATING SA PROBABILITY DISTRIBUTIONS: GENERALIZED REPLICATOR DYNAMICS

In this section, to simplify the denotations, we replace $\mathbf{p}^S$ as $\mathbf{x}$ and $\mathbf{p}^G$ as $\mathbf{y}$. In the case of symmetric

matrix, let $\mathbf{p}^S = \mathbf{p}^G = \mathbf{x}$.

As proved in [9], in the case that W is symmetric, replicator dynamics is able to approximate the maximization problem of $P(\mathbf{x}, \mathbf{x}, W)$ in Eq.(1). In this section, we firstly introduce replicator dynamics and then propose a novel discrete dynamical system, which generalize replicator dynamics from symmetric matrix to general matrix. Therefore, this new dynamical system is called "*Generalized Replicator Dynamics*" (shortly denoted as GRD). As GRD is developed based on all the evolutionary concepts of replicator dynamics, e.g., natural selection model, GRD can efficiently approximate the maximization problem of $P(\mathbf{x}, \mathbf{y}, W)$ in Eq.(1). Like replicator dynamics, we also provide the concrete proof to guarantee the optimization ability of GRD.

4.1. Replicator Dynamics

Replicator dynamics is one of the population dynamical methods, which is also a kind of discrete dynamical system. It was first introduced and studied in evolutionary game theory to model the evolution of animal behavior [10]. Motivated by population evolution, the idea of replicator dynamics has been independently studied in many fields, such as population genetics [4], mathematical ecology [3], computer vision [13]. Replicator dynamics is based on the classical *selection model* by studying the effect of selection upon a population. The differential viabilities of the genotypes are the key of selection.

Consider a single chromosomal locus with n alleles $A_1, \ldots, A_n$. Let $x_1^{(t)}, \ldots, x_n^{(t)}$ denote the gene frequencies at the mating stage in the parental generation (the t-th generation). The assumption of random mating leads to $x_i^{(t)} x_j^{(t)}$ for the probability that a zygote carries the gene pair (A_i, A_j). Let w_{ij} be the probability that an (A_i, A_j)-individual survives to adult age. Since the gene pairs (A_i, A_j) and (A_j, A_i) belong to the same genotype, the selective value $w_{ij} \geqslant 0$ and $w_{ij} = w_{ji}$. The selection matrix $W = (w_{ij})_{n \times n}$ is therefore symmetric.

If N is the number of zygotes in the new generation, the $(t+1)$-th generation, then $x_i^{(t)} x_j^{(t)} N$ of them carry the gene pair (A_i, A_j) of which $w_{ij} x_i^{(t)} x_j^{(t)} N$ survive to adulthood. Therefore, the total number of individuals reaching the mating stage is $\sum_{r,s=1}^{n} w_{rs} x_r^{(t)} x_s^{(t)} N$. Let f_{ij} denote the frequency of the gene pair (A_i, A_j) in the adult stage of the $(t+1)$-th generation, we can obtain,

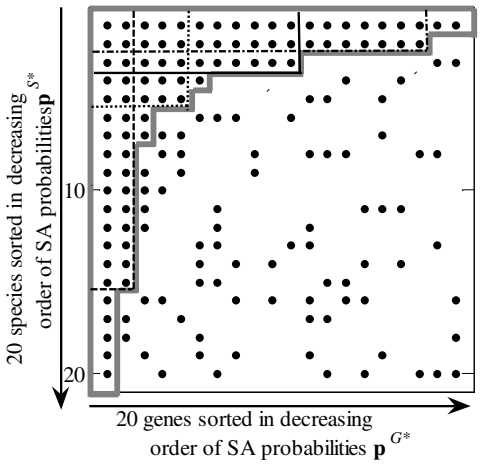


Fig. 3. Ladder structure in a small simulated species-genes matrix, a miniature of species-genes data. Its SA probability distributions of species and genes follow the power law distribution with $\alpha = 1$ and $\beta = 1$. A black dot indicates a non-zero value in both matrices. It can be seen that the ladder structure (framed by solid thick grey lines) exists in the simulated species-genes data. And bicliques (framed by lines with different types) in this structure overlap each other. The 10th specie (4 sequences available) and 20th specie (8 sequences available) are labeled in the left part of the figure.

$$f_{ij} = \frac{w_{ij} x_i^{(t)} x_j^{(t)} N}{\sum_{r,s=1}^{n} w_{rs} x_r^{(t)} x_s^{(t)} N} \tag{2}$$

Since $x_i^{(t+1)}$ is the frequency of the allele A_i in the adult stage of the $(t+1)$-th generation, we have $x_i^{(t+1)} = \sum_{j=1}^{n} f_{ij}$. This leads to the relation

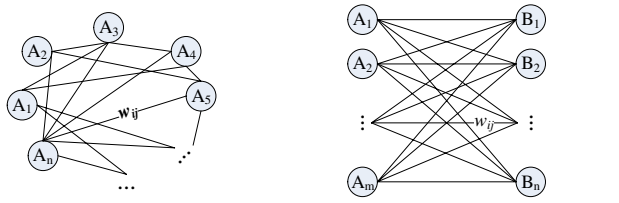
$$x_i^{(t+1)} = x_i^{(t)} \frac{\sum_{j=1}^{n} w_{ij} x_j^{(t)}}{\sum_{r,s=1}^{n} w_{rs} x_r^{(t)} x_s^{(t)}} \qquad i = 1, \ldots, n \tag{3}$$

Eq.(3) is the *selection model*. It can be rewritten in the matrix form as follows,

$$x_i^{(t+1)} = x_i^{(t)} \frac{(W\mathbf{x}^{(t)})_i}{\mathbf{x}^{(t)T} W \mathbf{x}^{(t)}} \qquad i = 1, 2, \ldots, n \tag{4}$$

where $(W\mathbf{x}^{(t)})_i$ denotes the i-th component of the vector $W\mathbf{x}^{(t)}$, and the state of the gene pool of the t-th generation is given by the vector $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_n^{(t)})^T$ of gene frequencies. $\mathbf{x}^{(t)}$ has non-negative components summing up to one, and belongs to the simplex Δ_+^n. Eq.(4) describes the ac-

tion of selection from one generation to the next, and therefore the map sending $\mathbf{x}^{(t)}$ to $\mathbf{x}^{(t+1)}$ defines a discrete dynamical system on the space Δ_+^n, called *Replicator Dynamics*.



(a) replicator dynamics (b) generalized replicator dynamics

Fig. 4. Alleles A_i or B_j as vertices and their mating survival probabilities w_{ij} as edge weights in replicator dynamics and generalized replicator dynamics.

Definition 4.1 (Replicator Dynamics). *Let $W_{n\times n}$ be a non-negative symmetric matrix. Given the vector $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_n^{(t)})^T \in \Delta_+^n$ being the status of the system in the t-th iteration, we define the dynamical system as Eq.(4).*

Since the selection model from evolutionary biology defines a discrete dynamical system *replicator dynamics*, we are interested in its stationary states and the optimization ability. Before that, we first introduce the average fitness of the population.

Definition 4.2. (Average Fitness of Population in Selection Model). Given $x_i^{(t)}x_j^{(t)}$ the frequency of the zygote of (A_i, A_j) and the selective value w_{ij} the probability that it survives to adult age, we define $\sum_{i,j=1}^n w_{ij}x_i^{(t)}x_j^{(t)}$ is the average fitness (or average selective value) of the population in the (t)-th generation. The average fitness can be written in the matrix form as $P(\mathbf{x}^{(t)}, \mathbf{x}^{(t)}, W) = \mathbf{x}^{(t)T}W\mathbf{x}^{(t)}$.

The fundamental theorem of natural selection tells us that under selection model, the average fitness increases from generation to generation. Refer to [9, 10] for detailed proof of this theorem.

Theorem 4.1. (Fundamental Theorem of Natural Selection by Replicator Dynamics). *For the replicator dynamics given by Eq.(4), the average fitness $P(\mathbf{x}^{(t)}, \mathbf{x}^{(t)}, W)$ increases with the generation t increasing in the sense that*

$$P(\mathbf{x}^{(t+1)}, \mathbf{x}^{(t+1)}, W) \geqslant P(\mathbf{x}^{(t)}, \mathbf{x}^{(t)}, W) \qquad (5)$$

with equality if and only if $\mathbf{x}^{(t)}$ is an equilibrium point $\mathbf{x}^*$.

4.2. Generalized Replicator Dynamics

The selection model above is based on the selection matrix $W_{n\times n}$ that describes the survival probability of the zygotes of any two alleles (A_i, A_j). Therefore, W is symmetric and the adjacency matrix of a weighted graph $G(A, W)$, whose vertex set A is alleles and edge weight is w_{ij} in W. This weighted graph is shown in Fig.4(a). In this section, we generalize the replicator dynamics to a more general selection matrix $W_{m\times n}$ that denotes the probability of the zygotes of any two alleles (A_i, B_j) from allele types A and B. We suppose that there are two types (or sets) of alleles $A = \{A_1, \ldots, A_m\}$ and $B = \{B_1, \ldots, B_n\}$. There are restrictions of mating in these two types of alleles: the mating can only happen between different types of alleles. For example, the allele A_i can mate with any B-type allele B_j, but always fail with any other A-type allele. Therefore, the selection matrix $W_{m\times n}$ and two sets of alleles A and B forms a bipartite graph as shown in Fig.4(b).

Let $x_1^{(t)}, \ldots, x_m^{(t)}$ denote the gene frequencies of A-type alleles $A_1, \ldots, A_m$, and $y_1^{(t)}, \ldots, y_n^{(t)}$ the gene frequencies of B-type alleles $B_1, \ldots, B_n$, at the mating stage in the parental generation (the t-th generation). The assumption of random mating leads to $x_i^{(t)}y_j^{(t)}$ for the probability that a zygote carries the gene pair (A_i, B_j). If N is the number of zygotes in the new generation, the $(t+1)$-th generation, then $x_i^{(t)}y_j^{(t)}N$ of them carry the gene pair (A_i, B_j) of which $w_{ij}x_i^{(t)}y_j^{(t)}N$ survive to adulthood. Therefore, the total number of individuals reaching the mating stage is $\sum_{r=1}^m \sum_{s=1}^n w_{rs}x_r^{(t)}y_s^{(t)}N$. Let f_{ij} denote the frequency of the gene pair (A_i, B_j) in the adult stage of the $(t+1)$-th generation, we can obtain,

$$f_{ij} = \frac{w_{ij}x_i^{(t)}y_j^{(t)}N}{\sum_{r=1}^m \sum_{s=1}^n w_{rs}x_r^{(t)}y_s^{(t)}N} \qquad (6)$$

Since $x_i^{(t+1)}$ is the frequency of the allele A_i in the adult stage of the $(t+1)$-th generation, we have $x_i^{(t+1)} = \sum_{j=1}^m f_{ij}$. This leads to the relation

$$x_i^{(t+1)} = x_i^{(t)} \frac{\sum_{j=1}^n w_{ij}y_j^{(t)}}{\sum_{r=1}^m \sum_{s=1}^n w_{rs}x_r^{(t)}y_s^{(t)}} \quad i = 1, \ldots, m$$

It can be rewritten in the matrix form as follows,

$$x_i^{(t+1)} = x_i^{(t)} \frac{(W\mathbf{y}^{(t)})_i}{\mathbf{x}^{(t)T} W \mathbf{y}^{(t)}} \qquad i = 1, 2, \ldots, m \quad (7)$$

For B-type alleles, since $y_j^{(t+1)}$ is the frequency of the allele B_j in the adult stage of the $(t+1)$-th generation, we have $y_j^{(t+1)} = \sum_{i=1}^{m} f'_{ij}$, where f'_{ij} is computed according to Eq.(6) by substituting $x_i^{(t)}$ with $x_i^{(t+1)}$. This leads to the relation

$$y_j^{(t+1)} = y_j^{(t)} \frac{\sum_{i=1}^{m} w_{ij} x_i^{(t+1)}}{\sum_{r=1}^{m} \sum_{s=1}^{n} w_{rs} x_r^{(t+1)} y_s^{(t)}} \quad j = 1, \ldots, n$$

Its matrix form is,

$$y_j^{(t+1)} = y_j^{(t)} \frac{(W^T \mathbf{x}^{(t+1)})_j}{\mathbf{y}^{(t)T} W^T \mathbf{x}^{(t+1)}} \qquad j = 1, 2, \ldots, n \quad (8)$$

The state of the gene pool of the t-th generation is given by the vector $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_m^{(t)})^T$ of gene frequencies in A-type alleles and the vector $\mathbf{y}^{(t)} = (y_1^{(t)}, \ldots, y_n^{(t)})^T$ of gene frequencies in B-type alleles. $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ have non-negative components summing up to one, and belong to the simplex Δ_+^m and Δ_+^n respectively. Eq.(7) and Eq.(8) are the *generalized selection model* for two types of alleles A and B. It describes the action of selection between two types of alleles from one generation to the next, and therefore the map sending $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ to $\mathbf{x}^{(t+1)}$ $\mathbf{y}^{(t+1)}$ defines a discrete dynamical system on the spaces Δ_+^m and Δ_+^n, called *Generalized Replicator Dynamics* (GRD).

Definition 4.3. (Generalized Replicator Dynamics). Let $W_{m\times n}$ be a non-negative matrix. Given the vector $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_m^{(t)})^T \in \Delta_+^m$ and the vector $\mathbf{y}^{(t)} = (y_1^{(t)}, \ldots, y_n^{(t)})^T \in \Delta_+^n$ being the status of the system in the t-th iteration, we define the discrete dynamical system as Eq.(7) and Eq.(8).

Correspondingly, we studied the the fixed points and optimization ability of the generalized replicator dynamics. Next the average fitness of the population and the fundamental theorem of natural selection in the generalized selection model are given.

Definition 4.4. (Average Fitness of Population in Generalized Selection Model). Given $x_i^{(t)} y_j^{(t)}$ the frequency of the zygote of (A_i, B_j) and the selective value w_{ij} the probability that it survives to adult age, we define $\sum_{i=1}^{m} \sum_{j=1}^{n} w_{ij} x_i^{(t)} y_j^{(t)}$ is the average fitness (or average selective value) of the population in the (t)-th generation. The average fitness in the matrix form is $P(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, W) = \mathbf{x}^{(t)T} W \mathbf{y}^{(t)} = \mathbf{y}^{(t)T} W^T \mathbf{x}^{(t)}$ and therefore the same as the form of accumulated probability function introduced in Section 3 of a bipartite graph $\mathcal{G}(A, B, W)$, where A and B are two sets of alleles representing the vertices.

Theorem 4.2. (Fundamental Theorem of Natural Selection by Generalized Replicator Dynamics). *For the generalized replicator dynamics given by Eq.(7) and Eq.(8), the average fitness $P(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, W)$ increases with the generation t increasing in the sense that*

$$P(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}, W) \geqslant P(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, W) \quad (9)$$

with equality if and only if $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ are two equilibrium points $\mathbf{x}^*$ and $\mathbf{y}^*$ respectively.

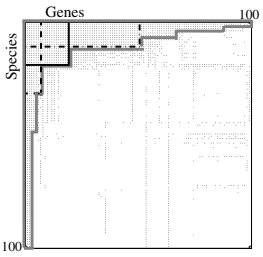
Proof. See Appendix A. □

If let W be symmetric, $\mathbf{x}$ and $\mathbf{y}$ are associated with the same set of vertices and thus equal to each other. Hence Eq.(7) and Eq.(8) are reduced to Eq.(4) and therefore the replicator dynamics become a special instance of the generalized replicator dynamics. In practice, the iteration of about 50 is enough for the generalized replicator dynamics to get converged. Therefore, its computational complexity is $\mathbf{O}(k(2h+m+n))$, where k is the number of iterations, h, m and n are the number of non-zeros, numbers of rows and columns in W respectively. If ignoring k, the final complexity is $\mathbf{O}(2h + m + n)$. Therefore, the generalized replicator dynamics is very efficient.

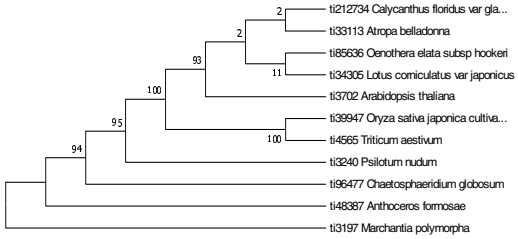
5. EMPIRICAL STUDY

To test if GRD is able to estimate the sequence-availability distributions for discovering the target pattern we proposed – *ladder structure*, we apply GRD to two data sets collected from GenBank and Swiss-Prot [e] respectively, which were published in [5]. Their species-genes data matrices are shown in Fig. 2(d) and 2(e). To validate the effectiveness of phylogenetic inference of *ladder structure*, bicliques

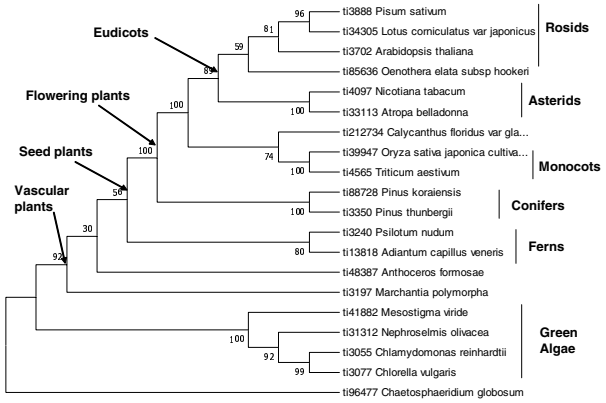
[e] both are available at http://ginger.ucdavis.edu/sandlab/www_data



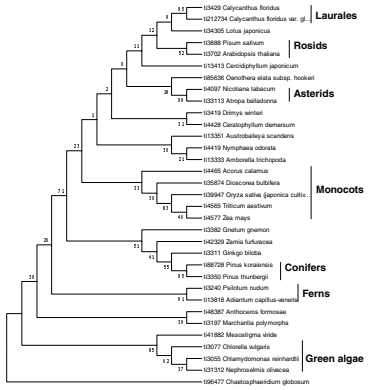
(a) the left-top corner of the reordered species-genes matrix.



(b) biclique 11×51, frame with dash-dot lines in (a)



(c) biclique 20×21, frame with solid lines in (a)

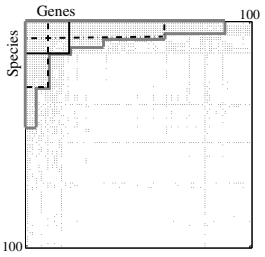


(d) biclique 33×5, frame with dashed lines in (a)

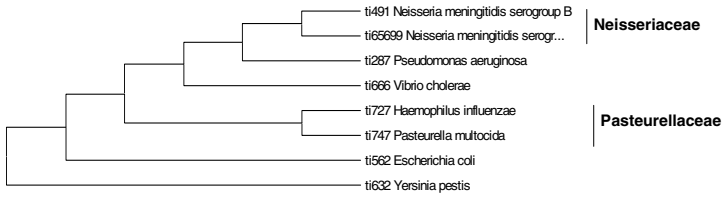
Fig. 5. GenBank: the phylogenetic trees of selected bicliques obtained from the submatrix 100×100 computed by the generalized replicator. The numbers on the branches of pylogenetic trees are bootstrap support. A black dot in (a) indicates a non-zero value in the matrix. The ladder structure (framed by solid thick grey lines) in (a) can be clearly seen.

in this overlapping structure are manually selected for investigation. This process has been illustrated in Fig. 1. In detail, given a biclique, following steps described in [5], sequences in bicliques are firstly concatenated and aligned using CLUSTALW with default options. Protein parsimony is used to construct trees, then bootstrap analysis (500 replicates) is applied to assess the reliability of trees, and finally the consensus tree is the eventual output. We implemented generalized replicator dynamics in MATLAB and all the experiments are performed in the computer system with Pentium 4 CPU 1.80GHZ, 512MB of RAM.

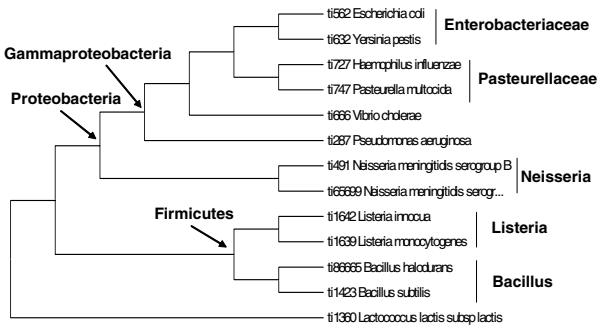
GenBank data is extracted from GenBank database. It contains 16,348 species and 59,144 genes. According to the result published in [5], there are 5587 bicliques with at least four species and two genes in the data. However, in this published result, the most distinct overlapping relationship (i.e., ladder structure) of bicliques among the 5587 bicliques is not revealed. It took GRD less than 1 second to discover the distinct ladder structure from GenBank data, while it took the biclique enumeration algorithm more than 900 seconds to find bicliques. After obtaining estimated SA distributions of species and genes, i.e., $\mathbf{p}^{S*}$ and $\mathbf{p}^{G*}$, we reorder the species-genes data matrix by the decreasing order of $\mathbf{p}^{S*}$ and $\mathbf{p}^{G*}$. According to our analysis in Section 2, the most distinct ladder structure is collected to the left-top corner of the reordered matrix. Therefore, as the original matrix is very large, we only show the first 100 species and 100 genes in the left-top corner of the reordered matrix in Fig.5(a). From this figure, a distinct ladder structure can be clearly seen and a lot of bicliques overlap in this ladder structure. Among these overlapping bicliques, we select three types of bicliques with different sizes: few species and many genes, balanced numbers of species and genes, and many species and few genes. They are 11×51, 20×21 and 33×5, and are framed in different types of dotted lines in Fig. 5(a). Their corresponding phylogenetic trees are also shown in Fig. 5(b–d). When



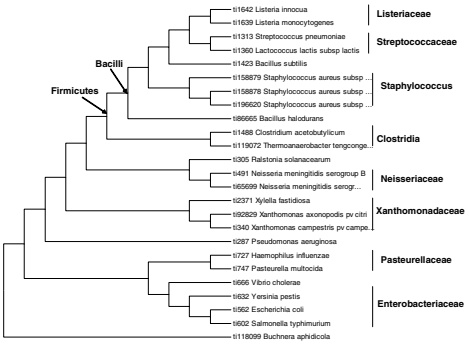
(a) the left-top corner of the reordered species-genes matrix.



(b) biclique 8×62, frame with dash-dot lines in (a)



(c) biclique 13×17, frame with solid lines in (a)



(d) biclique 25×7, frame with dashed lines in (a)

Fig. 6. SwissProt: the phylogenetic trees of selected bicliques obtained from the submatrix 100×100 computed by the generalized replicator dynamics. The numbers on the branches of pylogenetic trees are bootstrap support. A black dot in (a) indicates a non-zero value in the matrix. The ladder structure (framed by solid thick grey lines) in (a) can be clearly seen.

investigating these three trees, we found that the phylogenetic tree in Fig. 5(c) provides strong support for major clades within green plants, such as eudicots, flowering plants, seed plants, and vascular plants. In contrast, the trees in Fig.5(b) and 5(d) are not so informative. Phylogenetic trees from other bicliques with balanced numbers of species and genes in the same ladder structure also have similar results as shown in Fig. 5(c). We found that the relative positions of different organisms on these trees are not affected. In other words, organisms within the same genus are closely clustered together, e.g. Rosids, Asterids, Monocots, Conifers, Ferns and Green algae. This indicates that bicliques from the distinct ladder structure keep the stable inference of phylogenetic trees. This result shows that the distinct ladder structure is easy to discover and useful for phylogenetic inference. Furthermore, our model can not only locate many overlapping bicliques efficiently and effectively, but also reveal that these overlapping bicliques keep similar and stable phylogenetic structure, which provides more useful information to phylogeneticists for comparing and evaluating phylogenetic trees from species-genes data. This is not what the biclique enumeration method can get.

SwissProt data is extracted from Swiss-Prot database. It contains 7449 species and 64,712 genes. In this data set, we obtained similar results as those of Genbank data. They are shown in Fig. 6. Like GenBank data, the ladder structure can also be discovered in SwissProt data as shown in Fig. 6(a). Similarly, three bicliques are easily selected for building phylogenetic trees. These trees are presented in Fig. 6(b–d). The results in SwissProt data further verify the conclusions and efficiencies of our model.

6. CONCLUSIONS AND FUTURE WORKS

To better infer the evolutionary history of species, we build a model to understand and analyze species-genes data, also called sequence availability data. As previous works on species-genes data can provide only qualitative and rough view of this type of data, in this paper, we built a model to analyze

it in a quantitative way. Through this model, two conclusions are obtained: (1) It is the skewness of the sequence-availability probability distributions of species and genes that contribute to the sparseness and skewness real-world species-genes data. Further, the sequence-availability probability distributions of species and genes follow power law. (2) By estimating sequence-availability probability distributions of species and genes in real data, a distinct ladder structure is discovered, that is an overlapping structure of bicliques. To estimate sequence-availability probability distributions of species and genes in real data for finding the distinct ladder structure, we proposed a novel evolutionary dynamical system, called "generalized replicator dynamics", that is generalized from a popular biological system replicator dynamics. It is based on the fundamental theorem of natural selection and it can converge and approximate the solution of the maximization problem we formulated for the model estimation. We have conducted experiments on two species-genes data sets and the results have shown the effectiveness of our model in understanding and analyzing species-genes data for efficient phylogenetic inference.

There are a number of venues in the future works. Because there are not only one ladder structure in the real data, algorithms based on the model need to be developed to find more other ladder structures. Besides the aspect of algorithms, more experiments of our model in other species-genes data are needed for observing more properties and characteristics of the data for effective phylogenetic inference.

References

1. Alexe G., Alexe S., Crama Y., Foldes S., Hammer P.L., Simeone B. Consensus algorithms for the generation of all maximal bicliques. *Discrete Applied Mathematics* 2004; **145(1):** 11–21.
2. Bapteste E., Brinkmann H., Lee J.A., Moore D.V., Sensen C.W., Gordon P., Duruflé L., Gaasterland T., Lopez P., Müller M., Philippe H. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc. Natl. Acad. Sci. USA* 2002; **99(3):** 1414–1419.
3. Baum L. E., Eagon J. A. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Amer. Math. Soc.* 1967; **73:** 360–363.
4. Crow J. F., Kimura M. *An Introduction to Population Genetics Theory*. Harper & Row, New York, 1970.
5. Driskell A.C., Ańe C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. Prospects for building the tree of life from large sequence databases. *Science* 2004; **306:** 1172–1174.
6. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods in Enzymology* 1996; **266:** 419–427.
7. Felsenstein J. *Inferring Phylogenies*. Sinauer Press, 2003.
8. Hershkovitz M.A., Leipe D.D. *Bioinformatics: A practical guide to the analysis of genes and proteins*, chapter Phylogenetic analysis, pages 189–230. Wiley Interscience, New York, 1998.
9. Hofbauer J., Sigmund K. *The Theory of Evolution and Dynamical Systems*. Cambridge University Press, 1988.
10. Hofbauer J., Sigmund K. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.
11. Murphy W.J., Eizirik E., Johnson W.E., Zhang Y.P., Ryder O.A., O'Brien S.J. Molecular phylogenetics and the origins of placental mammals. *Nature* 2001; **409:** 614–618.
12. Page R.D.M., Holmes E.C. *Molecular Evolution: a Phylogenetic Approach*. Blackwell, 1998.
13. Pelillo M. The dynamics of nonlinear relaxation labeling processes. *J. Math. Imaging Vision* 1997; **7(4):** 309C323.
14. Qiu Y.L., Bernasconi-Quadroni F., Soltis D.E., Soltis P.S., Zanis M., Zimmer E.A., Chen Z., Savolainen V., Chase M.W. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 1999; **402(6760):** 404–407.
15. Russo C.A.M., Takezaki N., Nei M. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.* 1996; **13(3):** 525–536.
16. Sanderson M.J., Driskell A.C. The challenge of constructing large phylogenetic trees. *TRENDS in Plant Science* 2003; **8(8):** 374–379.
17. Sanderson M.J., Driskell A.C., Ree R.H., Eulenstein O., Langley S. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Mol. Biol. Evol.* 2003; **20(7):** 1036–1042.
18. Swofford D.L., Olsen G.J., Waddell P.J., Hillis D.M. *Molecular Systematics*, chapter Phylogenetic inference, pages 407–514. Sinauer Associates, Sunderland, Massachusetts, 2nd edition, 1996.
19. Wiens J.J. Missing data and the design of phylogenetic analyses. *Journal of Biomedical Informatics* 2006; **39(1):** 34–42.

APPENDIX A: PROOF OF THEOREM 4.2

For simplicity, Eq.(7) and Eq. (8) are rewritten as,

$$x'_i = x_i \frac{(W\mathbf{y})_i}{\mathbf{x}^T W \mathbf{y}} \tag{10}$$

$$y'_i = y_i \frac{(W^T\mathbf{x}')_i}{\mathbf{y}^T W^T \mathbf{x}'} \tag{11}$$

where $\mathbf{x}$ and $\mathbf{x}'$ represent $\mathbf{x}^{(t)}$ and $\mathbf{x}^{(t+1)}$, and similarly for $\mathbf{y}$ and $\mathbf{y}'$. Correspondingly, Eq.(9) is rewritten as,

$$P(\mathbf{x}', \mathbf{y}', W) \geqslant P(\mathbf{x}, \mathbf{y}, W) \tag{12}$$

It is clearly seen that $P(\mathbf{x}, \mathbf{y}, W) = \mathbf{x}^T W \mathbf{y} = \mathbf{y}^T W^T \mathbf{x} = \sum_{ij} x_i w_{ij} y_j$. We will prove the following two inequalities step by step,

$$\mathbf{x}'^T W \mathbf{y}' \geqslant \mathbf{x}'^T W \mathbf{y} \tag{13}$$

$$\mathbf{x}'^T W \mathbf{y} \geqslant \mathbf{x}^T W \mathbf{y} \tag{14}$$

Proof of Inequality (13)

Since we assume $\mathbf{x}'^T W \mathbf{y} > 0$, we have to show that

$$(\mathbf{x}'^T W \mathbf{y})(\mathbf{x}'^T W \mathbf{y}') \geqslant (\mathbf{x}'^T W \mathbf{y})^2 \tag{15}$$

Clearly,

$$(\mathbf{x}'^T W \mathbf{y})(\mathbf{x}'^T W \mathbf{y}') = (\mathbf{x}'^T W \mathbf{y}) \sum_{ij} x'_i w_{ij} y'_j \tag{16}$$

On replacing y'_j by the expression in Eq. (11) we obtain

$$\begin{aligned}
(\mathbf{x}'^T W \mathbf{y})(\mathbf{x}'^T W \mathbf{y}') &= (\mathbf{x}'^T W \mathbf{y}) \sum_{ij} x'_i w_{ij} (y_j \frac{(W^T\mathbf{x}')_j}{\mathbf{y}^T W^T \mathbf{x}'}) \\
&= \frac{\mathbf{x}'^T W \mathbf{y}}{\mathbf{y}^T W^T \mathbf{x}'} \sum_{ij} x'_i w_{ij} y_j (W^T\mathbf{x}')_j \\
&= \sum_{ij} x'_i w_{ij} y_j (W^T\mathbf{x}')_j \\
&= \sum_{ij} x'_i w_{ij} y_j \sum_k w_{kj} x'_k
\end{aligned} \tag{17}$$

Here, we use the *inequality of Cauchy-Schwarz-Bunyakowski*

$$(\sum_{i=1}^n a_i b_i)^2 \leqslant (\sum_{i=1}^n a_i^2)(\sum_{i=1}^n b_i^2) \qquad \text{for all} \quad a_i, b_i \geqslant 0 \tag{18}$$

with equality iff there is some value c such that $\frac{a_j}{b_j} = c$ for all j. By (18), we obtain

$$\begin{aligned}
(\mathbf{x}'^T W \mathbf{y})^2 &= (\sum_{ij} x'_i w_{ij} y_j)^2 \\
&= \Big[\sum_j y_j (\sum_i x'_i w_{ij})\Big]^2 \\
&= \Big[\sum_j (\sqrt{y_j})(\sqrt{y_j} \sum_i x'_i w_{ij})\Big]^2 \\
&\leqslant \Big[\sum_j y_j\Big]\Big[\sum_j y_j (\sum_i x'_i w_{ij})^2\Big] \\
&= \sum_j y_j (\sum_i x'_i w_{ij})^2 \\
&= \sum_j y_j \sum_i x'_i w_{ij} (\sum_k x'_k w_{kj}) \\
&= \sum_{ij} y_j x'_i w_{ij} (\sum_k x'_k w_{kj})
\end{aligned} \tag{19}$$

Combining Eq.(17) and Inequality (19), we prove Inequality (15). If $L_W(\mathbf{x}', \mathbf{y}') = L_W(\mathbf{x}', \mathbf{y})$ in Inequality (13), the last estimate must be an equality, i.e. there must be a value c such that $\frac{\sqrt{y_j} \sum_i x'_i w_{ij}}{\sqrt{y_j}} = (W^T\mathbf{x}')_j = c$ in Eq.(19) for all j. This means that $\mathbf{x}'$ is an equilibrium.

Proof of Inequality (14)

Similarly, we can follow the proof methodology of Inequality (13) to prove,

$$(\mathbf{x}^T W \mathbf{y})(\mathbf{x}'^T W \mathbf{y}) \geqslant (\mathbf{x}^T W \mathbf{y})^2 \tag{20}$$

If $L_W(\mathbf{x}', \mathbf{y}) = L_W(\mathbf{x}, \mathbf{y})$ in Inequality (14), there must be a value d such that $(W\mathbf{y})_j = d$ in Eq.(19) for all j. This means that $\mathbf{y}$ is an equilibrium.