RESOVING THE GENE TREE AND SPECIES TREE PROBLEM BY PHYLOGENETIC MINING

XIAOXU HAN

Department of Mathematics and Bioinformatics Program, Eastern Michigan University Ypsilanti, MI 48197, USA

The gene tree and species tree problem remains a central problem in phylogenomics. To overcome this problem, gene concatenation approaches have been used to combine a certain number of genes randomly from a set of widely distributed orthologous genes selected from genome data to conduct phylogenetic analysis. The random concatenation mechanism prevents us from the further investigations of the inner structures of the gene data set employed to infer the phylogenetic trees and locates the most phylogenetically informative genes. In this work, a phylogenomic mining approach is described to gain knowledge from a gene data set by clustering genes in the gene set through a self-organizing map (SOM) to explore the gene dataset inner structures. From this, the most phylogenetically informative gene set is created by picking the maximum entropy gene from each cluster to infer phylogenetic trees by phylogenetic analysis. Using the same data set, the phylogenetic mining approach performs better than the random gene concatenation approach.

1 Introduction

1.1. *Gene tree and species tree problem*

The gene tree/species tree problem is still an important problem in phylogenomics. A gene tree is a phylogenetic hypothesis constructed from one gene; it may not represent the true evolutionary history of the species [1]. On the other hand, a species tree reflects the species evolutionary history correctly, but it is generally unknown to investigators for a group of organisms. Incongruence among species trees and gene trees simply means that gene trees are not isomorphic with species trees. The incongruence occurs due to possible biological or analytical reasons in the phylogenetic reconstruction. The biological reasons include paralogy, lineage sorting and horizontal gene transfer [1,2,3]. The analytic reasons can be the data sampling bias and fit of the mathematical models in the phylogenetic analysis [4,5,6]. Both of them can lead to the artifacts in phylogeny. There are many excellent models and approaches to address the gene tree and species tree problem from different point of view based on these factors [7,8].

There are two most recently proposed approaches to solve the species tree and gene tree problem. The first approach is to use complete genome data in the phylogenetic inference [9,10]. This approach removes the possible gene trees and species tree problem since all information is incorporated in the tree inference by comparing gene contents or gene orders [9,10,11]. Although it shows strong potential, such approach suffers from the primitive mathematical model and temporal data scarcity problem [11]. For example, the main algorithm employed in the gene-order based phylogenomic reconstruction is break-point analysis, a method to minimize the number of breakpoints between

genomes, where a NP-hard problem has to be solved in the each iteration. The current genome data needed for genome based approaches are still far from abundant compared with general sequence data although more than 260 genomes already sequenced and thousands of genome sequencing is in progress [9]. Another approach is called gene concatenation [12]. Its main idea is to include more genes involved in gene tree reconstruction by randomly combining a set of widely distributed orthologous genes selected from genome data. Rokas *et.al.* [12] randomly concatenated genes from a set of selected 106 widely distributed orthologous genes for seven *Saccharomyces* species (*S. cerevisiae, S. paradoxus, S. mikatae, S. kudriavavzevii, S. bayanus, S. castellii, S. kluyveri*) and an outgroup *Candida albicans* in their experiment. They showed that the phylogenetic analysis results from maximum parsimony (MP) and maximum likelihood (ML) of the DNA and corresponding protein sequences were the species tree for at least twenty (an experimental number) gene combinations. The random gene combination strategy gives no consideration of gene functionality in the phylogeny. Such a random concatenation strategy may bring noise signals into the phylogenies because the 106 selected genes may not be a sampling bias free data set and different genes may have different evolution history. It is possible that the noise signals play negative roles in the phylogeny. The experimental gene concatenation number in phylogenetic inference has to be reobtained for different data sets each time through large scale simulations. Furthermore, there is little knowledge known about the genes involved in the phylogenetic reconstruction except the basic orthology. The ad-hoc strategy of the random concatenation method inhibits biologists from resolving the species tree and gene tree problem robustly in the phylogenomic era.

1.2. *Gene concatenation under Bayesian analysis*

We define terminology "gene convergence" and "tree posterior probability" in our phylogemomic analysis. A gene is defined as a convenegent gene or a "good" gene if its gene tree is the species tree under a phylogenetic reconstruction model R. Otherwise the gene is called a nonconvergent gene or a "bad gene". In the 106 gene set G used by Rokas *et. al.*(we also use the same gene set), there are 45 convergent genes and 61 nonconvergent genes under a Bayesian analysis with the GTR+Γ model [13, 14]; The tree posterior probability for a evolutionary tree inferred from Bayesian analysis $t_p = \prod_{i=1}^{|B_I|} b_p^i$ is defined as the multiplication of all posterior probabilities of its inferred branches, where the b_p^i is the posterior probability in the i^{th} inferred branch and B_I is the set of all inferred branches (the corresponding branches related to the outgroup taxon are excluded).

We conduct a random gene concatenation under Bayesian analysis as follows. We randomly concatenate the genes according to three cases: good gene concatenation, bad gene combination and random gene combination for total gene set. For each gene combination case, we generate 10 random data sets for Bayesian analysis under the GTR+Γ model and compute the expected tree posterior for the ten trials. We found that simulation results were congruent with the results from those of Rokas *et. al*'s. although

different phylogenetic analysis methods were employed. With the increase of the number of genes in the concatenation, we observe that the tree average posterior probability increases for each combination case (Figure 1). The final evolution tree inferred by Bayesian analysis is the species tree with maximum support (Figure 2).

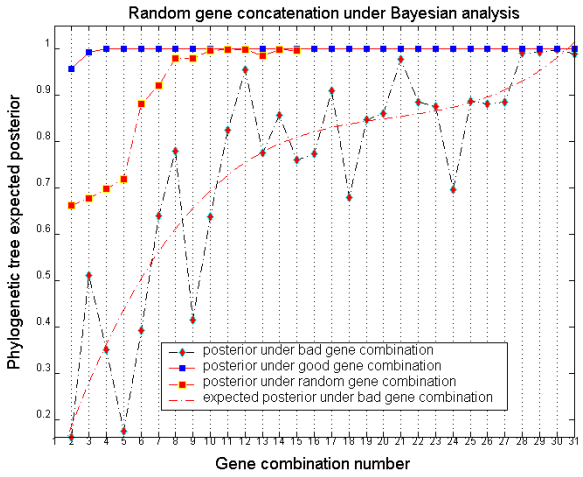

Figure 1. Random gene concatenation under Bayesian analysis.

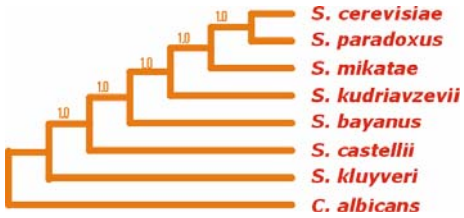

Figure 2. The species tree inferred by Bayesian analysis with the maximum support.

However, the performance from three types of combinations is quite different. To reach the best expected tree posterior probability, it requires at least 4 genes in the good gene combination case, at least 14 genes in the random gene combination case and at least 28 genes in the bad gene combination case, which is "the worst case" in this random gene combination method. In "the worst gene" combination case, there are still generally at least 10% gene combination sets whose gene trees are not the species tree if the gene combination sets have less than 28 genes. After computing the phylogenetic tree posterior probability standard deviations of gene combinations under bootstrap with 1000 bootstrap samples, we found that the bad gene combination had highest level oscillations and the good gene combination had lowest level oscillations in the plots of tree posterior probability standard deviations.

Actually, biologists don not know which genes are good or bad before knowing the species trees for a set of organisms. Thus, the worst gene combination case is unavoidable in the ad-hoc gene combination approach because investigators are not only blind to the inner structure of the gene set but also have little knowledge about which genes are more phylogenetically informative, not to mention the possible situation where

noisy data in the phylogenetic inference increase after gene combination.

1.3. *Overview of the methods and results*

In this paper, we develop a robust phylogenomic mining approach to address the gene tree/species tree problem. Self-organizing map (SOM) mining is employed to acquire knowledge from the gene set G using same data set as Rokas *et al.* [12] before the phylogenetic analysis starts. The genes in the gene set G are clustered into different sets according to their gene prototypes obtained from SOM mining. Then the maximum entropy gene from each cluster is selected to create a core gene set G_{core}. A phylogenetic tree will be inferred based on the core gene set. Our experimental results show the phylogenetic analysis (Maximum likelihood, Bayesian analysis) based on G_{core} always infers the desired species tree robustly. Compared with previous approaches, our phylogenomic mining approach can overcome the incongruence between the species tree and gene tree systematically. To our knowledge, this is the first work to integrate SOM mining and entropy theory in phylogenomics.

The paper is organized as follows. In the next section, we describe the basic idea of the phylogenomic mining. The third section describes the detailed phylogenomic mining method and related results. The fourth section presents gene entropy theory and its applications in phylogenetic inference. In addition to exploring the phylogenetic properties of the maximum entropy genes by Shimodaira-Hasegawa test, the Bayesian analysis is conducted to the core gene set created by picking maximum entropy genes from each gene cluster. The Bayesian phylogenetic analysis results are also compared with the random and good gene concatenation cases. Finally, we discuss our results and directions for future work.

2 Phylogenomic Mining: Knowledge Discovery from a Gene Set by Self-Organizing Map (SOM) Learning

The random gene concatenation method treats all genes uniformly in the phylogeny. The basic idea of the phylogenomic mining is to gain some knowledge from a gene set G by some unsupervised data mining algorithms before phylogenetic tree reconstruction. The gene set consists of aligned orthologous genes in our data set (the gene set actually can be rather huge, for example, a set of genomes.) The knowledge learning from gene set G is clustering genes in the gene set to gain insight into the possible structures of the gene set G, to identify some outliers, and to find possible phylogenetically informative genes. These phylogenetically informative genes may contain more evolutionary information based on a gene entropy measure; if so, they will be selected to create a core gene set G_{core}. The final phylogenetic tree will be reconstructed from the gene set carrying the most phylogenetic information. We will demonstrate that the final evolutionary tree from this selected set converges to the species tree with maximum support.

It is highly discouraged to cluster genes directly since different genes have different lengths. To cluster genes directly, Genes have to be encoded into uniform column

vectors. That is, zeros have to be filled into the encoded vectors for the short genes. The filled zeros will act as "missing data" in the clustering and it is not advisable to do clustering under such a condition [15]. Thus, we suggest an alternative way to cluster genes. The core idea of the phylogenomic mining is to cluster the prototypes of genes rather than genes themselves directly.

The prototype of a gene is a small dataset containing representative features of genes, compared with a gene (generally a high dimensional dataset). For example, the probability mass function of a gene $x = x_1 x_2 ... x_n$ on the R^2 space can be a gene prototype. Self-organizing map (SOM) is a traditional data mining approach and vector quantization model to map a high dimensional data set D to its representative prototype $W \in R^2 : SOM : D \to W \in R^2$. It takes competitive unsupervised learning (self-organization) to partition original data space into a set of corresponding representative prototypes. With the assumption of no prior knowledge about the data to be mined, the unsupervised learning in SOM is a process of feature extraction and data dimension reduction. SOM mining is widely used in gene expression data processing, vector quantization, visualization and commercial database mining [15,16,17].

3 The Phylogenomic Mining Method in Detail

The phylogenomic mining method to resolve the gene tree and species tree problem mainly consists of the following six steps: 1.Encoding the gene set G into a digital matrix D; 2. Conducting SOM mining for the data set D; 3.Computing the gene distribution on SOM plane P for each gene; 4.Clustering gene hierarchically; 5.Selecting maximum entropy genes from each cluster to build a core gene set; 6.Conducting phylogenetic analysis for the core gene set to infer gene trees.

The first step is to encode an input character matrix to its corresponding digital matrix D before phylogenomic mining. In our analysis, the input character matrix (the transposition of the general character matrix) is a 127026×8 matrix which consists of 106 genes of the eight yeast species: Scer, Spar, Smik, Skud, Sbay, Scas, Sklu, Calb. Each column represents a taxon, each row represents a site and a gene crosses certain number of rows according to its length. The four orthogonal bases are used to encode 'A', 'T',' C', 'G' respectively: A=(1,0,0,0)′, T=(0,1,0,0)′, C=(0,0,1,0)′, G=(0,0,0,1)′. Missing nucleotides and gaps are encoded as a vector with four zeros entries. The input character matrix is encoded as a 508104×8 digit matrix D after such encoding processing.

The digit matrix D is then sent to a self-organization map (SOM) for mining in the second step. The SOM takes an input/output plane with 20×20 neurons and employs sequence learning algorithm. The reference vectors are initialized by the principal component analysis. The neighborhood kernel function used is a Gaussian function. The prototypes of species, which are also called the species patterns, can be obtained immediately after the SOM mining. It is just the final reference vector matrix W. The iteration process is time consuming since the complexity of each training epoch is $O(nmk)$ where the sample number $n = 508104$, species number $m = 8$ and the number of neurons $k = 400$. In our analysis, the SOM mining takes more than 8 hours time to get the species prototype after more than 2000 epochs in a Pentium 4 with a 2.1 GHz CPU.

The third step in our method is to compute the prototypes of genes which are the gene distributions on the SOM plane. The gene prototypes are represented in a matrix $W_1(l \times k)$ in our computing, where l is the number of genes involved in the mining ($l = 106$) and k is the number of neurons on the SOM plane P.

For a multi-species gene $x = x_1 x_2 ... x_n$, its probability mass function on the SOM plane P is its prototype $y = y_1 y_2 ... y_k$ after SOM mining, where k is the number of neurons on the SOM plane. The prototype of a gene is a set of representative features extracted from the dataset x. It will follow the statistical property of the gene x and can be considered as an approximation to the original probability mass function $p(x)$ [18].

The gene distribution on the SOM plane P can be computed as follows. For a gene sample x_i (a site/character), there exists a best match unit j ($j = 1,2,...k$) whose reference vector is most similar to x_i in the Euclidean distance measure. Then for each neuron j on the SOM plane, there exist a sample set $s_j = \{x_{i_1}, x_{i_2}, ... x_{i_j}\}$, each of its entry acknowledges the neuron j as its best match unit (BMU). That is,

$$s_j = \{x \mid \arg\min_i \| x - w_i \| = j, i = 1,2,...n\} \quad (1)$$

The size of s_j : $| s_j | = h_j$ stands for the number of gene samples of the gene x which are closest to the reference vector in the neuron j in the Euclidean distance; that is, there are h_j gene samples hitting the j^{th} map unit on the SOM plane. So for each gene x in the gene set G, there exists a corresponding hit number sequence $h_x = h_1 h_2 ... h_k$ recording the projection of the gene on the SOM plane P.

Thus, the prototype of the gene x is a sequence of hit rate: $y = y_1 y_2 ... y_k$, where y_i is the hit rate of the gene x on the i^{th} neuron on P: $y_i = h_i / \Sigma_{j=1}^{k} h_j$. The gene distribution on the SOM plane P extracts the representative features from each gene in a uniform format by tracing the projection of each gene on the SOM plane P. Compared with the raw genes, the prototypes of genes are more representative and make gene features and patterns comparable and visualizable. The gene clustering based on the prototypes can help us identify the clusters of genes sharing same features which are biological meaningful. From the visualizations of gene distributions on SOM plane (data not shown), we find that the distributions show interesting patterns: the genes with similar gene distributions are clustered close together on the SOM plane. This is actually the characteristic of the SOM learning, that is, data closer to each other in the high dimensional space are mapped to the neurons in R^2 close to each other topologically [18].

The fourth step in our phylogenomic mining clusters genes hierarchically based on the prototypes of genes. Although SOM itself is a fuzzy clustering algorithm, where similar data samples are mapped in the neighborhoods of BMUs, the data prototype still needs to use hierarchical or partitive clustering to explore the global similarity in the data set [19]. The gene clustering is conducted through the hierarchically clustering the prototypes of genes into natural grouped clusters in our analysis. The natural division of the prototypes is achieved by specifying the inconsistency coefficiency or cutoff number obtained by computing U-matrix for prototype vectors [19]. In our experiment, the 106 orthologous genes from eight species are clustered as 18 naturally grouped clusters. Gene

clustering helps us to discriminate between the genes before phylogenetic analysis to overcome the "blindness" in the phylogeny reconstruction where no consideration is given to relationships between genes. It is interesting to see that each cluster shares similar phylogenetic properties in addition to the fact that they have similar gene distributions on the SOM plane. For example, all genes in cluster *4, 5, 12, 15, 18* are nonconvergent ("bad") genes.

4 Exploring Phylogenetically Informative Genes by Gene Entropy

After the gene clustering is completed, a core gene set G_{core} will be built by selecting phylogenetically informative ("important") genes from each cluster before the phylogenetic analysis to infer the final tree. But which genes are phylogenetically informative genes in a cluster and how do we identify them? The initial thought is to employ principal component analysis. But it cannot return to the original gene space since the final data we need are symbolic data in the phylogenomic reconstruction. Thus, another method has to be found to measure the utility of genes in phylogenetic reconstruction. Recalling that a gene can be a set of patterns generated from an alphabet set $\Gamma = \{A,T,C,G\}$, we borrow entropy from information theory to measure the phylogenetic informative potentials of a multiple species gene. The informative genes will selected from each cluster according to gene entropy values to build a core gene set for further phylogenetic analysis.

We give the definition of gene entropy as follows. For a gene $x = x_1 x_2 ... x_n$ (which is a character matrix, a set of aligned sequences; each character x_i is a column in the character matrix), the gene entropy is defined as

$$h(x) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i)) \tag{2}$$

The gene entropy is equivalent to the block entropy definition in DNA entropy if the character matrix is converted to a corresponding one dimensional sequence and the block length is the length of a site. Although the entropy estimate can be conducted by block entropy estimate approaches theoretically, the estimate may not be satisfied since the block length is generally assumed a large number [20]. Since the gene distribution on the SOM plane is the frequency distribution on the SOM based on for each site, such gene prototype is perfectly fitted to estimate the gene entropy, where the gene probability mass function on the SOM plane is the distribution of each site in the gene set.

We predict that a gene with large entropy value will contain more phylogenetic information. Thus, a higher entropy gene is highly likely to be a phylogenetically informative gene, for example, a "good" gene (a gene whose gene tree is nearest to the species tree). To verify such hypothesis, we conduct the Shimodaira-Hasegawa test (SH test) [21] to compute the delta log likelihood and associative p-value for each gene tree with respect to the species tree after ML analysis for each gene. The SH-test is conducted under the GTR+Γ model with reestimated log likelihoods (RELL) approximation for 21 tree topologies for each gene. From the SH-test results, we can see 8/10 genes in the high entropy zone (HEZ) are convergent, (where gene entropy is not less than the sum of mean and standard derivation of all gene entropy in the gene set G). On the other hand,

there are 10/12 genes in the low entropy zone (LEZ) that are nonconvergent genes, (LEZ is a set of genes whose gene entropy is not greater than the difference of mean and standard derivation of all gene entropy in the gene set G). There are 31 genes among the 45 convergent genes in the 106 gene set that have entropy greater than the average entropy of the gene set. In sum, higher entropy genes are more likely to be "good" genes in phylogenetic reconstruction.

4.1 Maximum entropy gene concatenation

It is reasonable to combine maximum entropy genes from each cluster into a core gene set and conduct phylogenetic analysis for this core gene set, because a high entropy gene appears to be more likely to be phylogenetically informative. A maximum entropy gene is the "local maximum" gene whose entropy is the maximum for all genes in a particular cluster instead of among the total gene set. Compared with the general random gene concatenation, this phylogenomic mining based approach may be more systematic and robust to resolve the incongruence in the phylogeny because gene combinations are done based on an information measure after phylogenomic mining. There is no experimental gene number computing needed if this new approach is applied to different data set. On the other hand, maximum entropy gene selection from each cluster can remove potential "noise" signals contained in the non-convergent genes, which appear from our analyses, more likely to include low entropy genes. Furthermore, maximum entropy gene concatenation based on clustering prevents from the "data bias" problem if the total gene set obtained has over-representation of one or several types of genes due to data acquisition issues. This property may ameliorate over-support for some branches in the final consensus tree.

We conduct following three experiments to build a core gene set by selecting maximum entropy genes to infer phylogenetic trees. In experiment 1, we build the core gene set G_{core} by selecting a maximum entropy gene from each cluster among 18 clusters. The core gene set consists of following 18 maximum entropy genes selected from available gene clusters: {*YIL125W, YNL220W, YOL049W, YNL082W YDL195W, YPL169C, YDL116W, YJL085W, YDL126C, YMR186W, YOR158W,YPL210C, YJL087C, YNR008W, YLR253W, YIL109C, YFR044C, YGR285C*}. The mean entropy of the core gene set is 5.2325 bits. The gene tree inferred from the Bayesian analysis under GTR+Γ model is the species tree where the posterior probability of 1.0 is on each inferred branch (Figure 2).

In experiment 2, we can further cluster the total gene set G into an arbitrary number of clusters and pick the maximum entropy genes from each cluster. This approach answers the query: suppose the gene set G is clustered as $\#j$ clusters, which genes are the most informative to build a robust phylogenentic tree? If we treat the whole gene set as a cluster (*i.e.* no clustering work), the maximum entropy gene 18 (*YDL126C*) will be the only gene in the core gene set. The corresponding tree probability t_p for gene 18 is 1.0 after Bayesian analysis. Similarly, the core gene set consists of $\#j$ maximum entropy gene selected from $\#j$ clusters if the total gene set G is clustered into $\#j$ sets. We test all cases for $\#j$ from 1 to 14, the corresponding tree posterior probability for each case is 1.0 after Bayesian analysis. Such striking results suggest that our approach can be useful to infer the species tree systematically and robustly. However, such maximum entropy gene selection depends on the gene clustering. Randomly selecting several higher

entropy genes may not produce good results because more genes selected from a same cluster will make the sampling bias occurrence which may give "over support" for specific branch patterns.

In experiment 3, we compare phylogentically informative gene (maximum entropy gene selection from each cluster) concatenation under phylogenomic mining with random and good gene concatenations. Just as before, for each gene combination case, we generate 10 random data sets for Bayesian analysis under the GTR+Γ model and compute the expected tree posterior for the ten trials. We found our approach performance is even better than the good gene combination performance from the phylogenetic tree expected posterior probability analysis (Figure 3). Considering the species tree is actually unknown to investigators, we suggest that our method can provide a systematic solution to the gene tree and species tree problem. Moreover, it is well suitable to identify potential phylogenetically informative genes by phylogenomic mining in large genome databases.

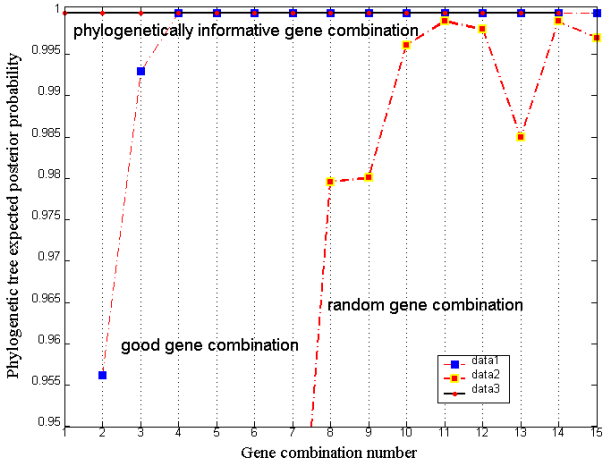

Figure 3. Comparing the performance of random and good gene concatenation cases with phylogenetically informative gene concatenation approaches.

5 Discussion and Future Work

In this paper, we provide a novel analytical solution to resolving the gene tree and species tree problem from phylogenomic mining point of view. Our results show that this approach is more robust than an ad-hoc gene concatenation approach. If generalizable, we also plan to take advantage of the powerful feature extract capability of SOM mining and entropy theory to address paralogy and orthology detection problem. The detection of paralogy and orthology is a key problem in phylogenomics but still in its naïve stage [3,6]. We expect the detection of the paralogous and orthologous genes can be conducted in their corresponding feature spaces and by means of entropy estimations. We are also interested in integrating our phylogenomic mining approach to the Bayesian analysis to explore more powerful tree reconstruction algorithms.

References

1. R. D. Page and E. Holmes. *Molecular evolution, a phylogenetics approach,* Blackwell Science, 1998.
2. W. Maddison. Gene trees in species trees. *Syst. Biol.,* 46:523-536, 1997.
3. J. Cotton. Analytical methods for detecting paralogy in molecular datasets, *Methods in Enzymology,* 395**:**700-24, 2005.
4. J. Huelsenbeck. Performance of phylogenetic methods in simulation. *Syst. Biol.,* 44:17–48, 1995.
5. Z. Yang, N. Goldman and A. Friday. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.*11:316–324, 1994.
6. K. Crandall and J. Buhay. Genomic databases and tree of life, *Science*, 306: 1144-1145, 2004.
7. B. Ma, M. Li, and L. Zhang. From Gene Trees to Species Trees. *SIAM Journal on Computing,* 30:729-752, 2000.
8. R. D. Page. Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Molecular Phylogenetics and Evolution*, 14:89-106, 2000.
9. F. Delsuc, H. Brinkmann and H. Philippe. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet.,* 6:361-375, 2005.
10. B. Snel, M. Huynen and B. Dutilh. Genome Trees and the Nature of Genome Evolution. *Annu. Rev. Microbiol.,* 59:191-209, 2005.
11. B. Moret, J. Tang, and T. Warnow. Reconstructing phylogenies from gene-content and gene-order data. *Mathematics of Evolution and Phylogeny,* Oxford Univ. Press, 321-352, 2005.
12. A. Rokas, B. Williams, N. King and S. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature,* 425:798-804, 2003.
13. J. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics,* 17:754-755, 2001.
14. J. Huelsenbeck and F. Ronquist. Bayesian Analysis of Molecular Evolution using MrBayes, http:// www.csit.fsu.edu/~ronquist/mrbayes/mrbayes_chapter.pdf**,** 2004.
15. M. Dunham. *Data mining introductory and advanced topics*, Prentice Hall, 2002.
16. S. Haykin. *Neural Networks: A Comprehensive Foundation, 2nd Edition.* Prentice-Hall, 1999.
17. J. Nikkila, P. Toronen, S. Kaski, J. Venna, E. Castren, G. Wong. Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks,* 15:953-966, 2002.
18. T. Kohonen. *Self-Organizing Maps, 3rd edition*, Berlin: Springer-Verlag, 2001.
19. J. Vesanto and E. Alhoniemi. Clustering of the Self-Organizing Map, *IEEE Transactions on Neural networks.* 11:586-600, 2000.
20. J. Lanctot, M. Li, and E. Yang. Estimating DNA sequence entropy, *Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms*. 409-418, 2000.
21. H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol.,* 16:1114-1116, 1999.