Statistical Applications in Genetics and Molecular Biology

Volume 3, *Issue* 1 2004 *Article* 31

Maximum Likelihood for Genome Phylogeny on Gene Content

Hongmei Zhang* Xun Gu†

*University of West Florida, hzhang@uwf.edu

†Iowa State University, xgu@iastate.edu

Maximum Likelihood for Genome Phylogeny on Gene Content*

Hongmei Zhang and Xun Gu

Abstract

With the rapid growth of entire genome data, reconstructing the phylogenetic relationship among different genomes has become a hot topic in comparative genomics. Maximum likelihood approach is one of the various approaches, and has been very successful. However, there is no reported study for any applications in the genome tree-making mainly due to the lack of an analytical form of a probability model and/or the complicated calculation burden. In this paper we studied the mathematical structure of the stochastic model of genome evolution, and then developed a simplified likelihood function for observing a specific phylogenetic pattern under four genome situation using gene content information. We use the maximum likelihood approach to identify phylogenetic trees. Simulation results indicate that the proposed method works well and can identify trees with a high correction rate. Real data application provides satisfied results. The approach developed in this paper can serve as the basis for reconstructing phylogenies of more than four genomes.

KEYWORDS: Gene Content, Maximum Likelihood, Phylogenetic Trees, Genome

---

*This study was supported by the NIH grant GM62118 to X. G.

Introduction

With the rapid growth of entire genome data, reconstructing the phylogenetic relationship among different genomes has become a hot topic in comparative genomics. Various approaches have been proposed. For instance, several authors (Snel et al. 1999; Huynen et al. 1999; Lin and Gerstein 2000; Korbel et al. 2002) estimated the fraction of shared genes for genome pairs, and transformed it to the genome distance matrix by some *ad hoc* distance measures. Other methods include the coefficient of co-occurrence of genomics (Natale et al. 2000) and the ratio of orthologs to the number of genes in the smaller genome (Clarke et al. 2002). In addition, various parsimony algorithms have also been used (e.g., Fitz-Gibbon and House 1999; House and Fitz-Gibbon 2002). It is true that almost no approaches are perfect. For example, trees reconstructed based on genome distances sometimes can be misleading (Huson and Steel 2004). The key criticism to the above genome approaches is the lack of clear-cut evolutionary model (Wolf et al. 2002). To resolve this problem, we (Gu and Zhang 2004) proposed a genome-distance method that is feasible in practice for genome-wide analysis. We have further demonstrated that, in order to estimate an additive genome distance, the typical gene content data, i.e., the presence or absence of gene families across genomes, is not sufficient. We overcome this difficulty by introducing the concept of extended gene content, that is, the status of a gene family in a given genome could be either absence, presence as single-copy, or presence as duplicates, which can be used to develop a distance-based method for phylogenetic inference.

The maximum-likelihood approach for tree reconstruction based on nucleotide or amino acid sequences has been very successful, and has several apparent advantages over other methods (Felsenstein 1981; Strimmer and Haeseler 1996), but there is no reported study for any applications in the genome tree-making. There are two difficulties. One is that the maximum likelihood approach requires an explicit probabilistic model for the genome evolution. The second is about the potential huge calculation burden at the genome-level. Gu (2000) developed a likelihood framework for phylogenetic gene-content analysis but it may require a huge amount of computational time.

In this paper, we concentrate on the four-genome case. We attempt to study the mathematical structure of a stochastic model of genome evolution, and develop a fast maximum likelihood algorithm used to infer phylogeny based on gene content information of genomes.

The birth-death model

Gu (2000) suggested a birth-death model assuming a varied average size of gene family for a long-term evolution. In one genome, let λ denote the evolutionary rate of gene proliferation, and μ be the evolutionary rate of gene loss. Conventionally, it is assumed that every member gene is subject to the same risk to be lost or duplicated. For a given gene family, let X_t be the number of member genes in a genome after t time units of evolution, and r_0 denote the number of member genes at the origin where $t = 0$. The transition probability of $X_t = k$ at time t is (Chiang 1968):

$$\begin{aligned} P(X_t = k|X_0 = r_0) &= \sum_{j=0}^{\min[r_0,k]} \frac{r_0!}{j!(r_0-j)!}\frac{(r_0+k-j-1)!}{(k-j)!(r_0-1)!} \\ &\quad \beta^{r_0-j}\alpha^{k-j}(1-\alpha-\beta)^j, \quad k \geq 1 \\ P(X_t = 0|X_0 = r_0) &= \beta^{r_0}, \end{aligned} \tag{1}$$

where the two parameters α and β are given by

$$\begin{aligned} \alpha &= \lambda\frac{1-e^{(\lambda-\mu)t}}{\mu-\lambda e^{(\lambda-\mu)t}} \\ \beta &= \mu\frac{1-e^{(\lambda-\mu)t}}{\mu-\lambda e^{(\lambda-\mu)t}}, \end{aligned} \tag{2}$$

respectively.

From (2) it is easy to see $\alpha/\beta = \lambda/\mu$. α and β are therefore called the proliferation and loss parameters, respectively. The expectation of a gene family size at time t is $r_0 e^{(\lambda-\mu)t}$ (Chiang, 1968), so $\alpha > \beta$ indicates, on average, the increase of a gene family size and *vice versa*.

The likelihood for a phylogenetic tree of four genomes

With four genomes, 15 possible rooted phylogenetic trees exist but these trees only encompass two different shapes as indicated in Figure 1. In the following we derive the likelihood function for the shape shown in Figure 1(a). The likelihood function for Figure 1 (b) can be derived in the same way (see Appendix 1 for the derivation).

For a given gene family, assume that there are X_0 member genes at $t = 0$ (the common ancestor), and r_A and r_B member genes at the two inner nodes,

respectively. We let $X_i, i = 1, 2, \cdots, 4$, be the number of member genes in genome i. We use $\alpha_i,\ i = 1, \cdots, 6$, to denote the gene proliferation parameter for each branch, and $\beta_i,\ i = 1, \cdots, 6$, to denote the gene loss parameter for each branch.

The number of member genes at the origin, X_0, is unknown. We assume X_0 follows a geometric distribution, i.e. $\pi(X_0 = r_0) = (1 - f)^{r_0} f$. This assumption follows the common sense that the probability of getting a large number of member genes at the origin is smaller than that of having a small number of member genes at the origin. As seen in the following derivation of the likelihood function, the geometric distribution also provides mathematical convenience. f is biologically explained as the probability of a gene family having one member gene. Its value is assumed to be known, and can be set as a commonly used value, or can be obtained from the size distribution of gene families if such information available.

Under the assumption of evolution independence between lineages, the marginal joint probability for the number of member genes of a gene family in the four genomes is given by

$$\begin{aligned} P(X_1, X_2, X_3, X_4) &= \sum_{r_0} \sum_{r_A} \sum_{r_B} \pi(r_0) P(r_A | r_0; \alpha_5, \beta_5) P(r_B | r_0; \alpha_6, \beta_6) \\ &\quad \times P(X_1 | r_A; \alpha_1, \beta_1) P(X_2 | r_A; \alpha_2, \beta_2) \\ &\quad \times P(X_3 | r_B; \alpha_3, \beta_3) P(X_4 | r_B; \alpha_4, \beta_4). \end{aligned} \tag{3}$$

The transition probabilities in (3) are calculated using (1).

If M gene families are present in the four genomes, the likelihood function for the parameters of $\alpha_i,$ and $\beta_i, i = 1, \cdots, 6$, is then given by

$$L(\alpha_1, \cdots, \alpha_6, \beta_1, \cdots, \beta_6 | \text{data}) = \prod_{m=1}^{M} P(X_{1m}, X_{2m}, X_{3m}, X_{4m}), \tag{4}$$

where $X_{im}, i = 1, \cdots, 4, m = 1, \cdots, M$, is the number of member genes for gene family m in genome i.

We notice that (3) can not be further simplified, and the three summations in (3) can give a heavy burden toward the calculation of (4). Also, since the cut-off value for member genes clustering may not be accurate enough, the calculated likelihood may not reflect the true likelihood. Consequently, the quality of the maximum likelihood estimates of the parameters can hardly be guaranteed. These two concerns, especially the calculation burden, implies that it is not efficient to use the original gene family frequency data to calculate the likelihood. Fortunately, as discussed below, the likelihood function can be simplified when we use gene contents.

Gene content data only need the extinction and existence information of a gene family instead of the number of member genes of each gene family. This type of data is commonly used in many primary biological researches. Gene content is a binary (1 or 0) index for a gene family in a genome. Let $Y = 1$ denote the situation that this genome maintains at least one member gene of the gene family after t time units of evolution, i.e., $X_t \geq 1$, and $Y = 0$ be the situation that all the member genes are lost, i.e. $X_t = 0$. Phylogenetic gene-content analysis only considers the presence ($Y = 1$) or absence ($Y = 0$) of a gene family. To distinguish between presence and absence of a gene family is much easier than to distinguish between having, say, 3 and 4 member genes, and thus more accurate.

From (1), the probabilities of $Y = 0$ is

$$P(Y = 0|X_0 = r_0) = P(X_t = 0|X_0 = r_0),$$

resulting in

$$P(Y = 0|X_0 = r_0) \quad = \quad \beta^{r_0}. \tag{5}$$

Clearly, we have $P(Y = 1|X_0 = r_0) = \sum_{k\geq 1} P(X_t = k|X_0 = r_0) = 1 - \beta^{r_0}$.

For the shape displayed in Figure 1(a), let Y_i be the gene content index of a gene family in genome i, $i = 1, \cdots, 4$. There are 16 possible gene content configurations in total for each gene family, i.e. $(Y_1, Y_2, Y_3, Y_4) = (1, 1, 1, 1)$, $(1, 1, 1, 0)$, ..., $(0, 0, 0, 0)$.

The joint probability of Y_1, Y_2, Y_3, and Y_4 can be easily derived using (3) and (5), which is

$$\begin{aligned} P(Y_1, Y_2, Y_3, Y_4) \quad &= \quad \sum_{r_0=1}^{\infty} \sum_{r_A=0}^{\infty} \sum_{r_B=0}^{\infty} \pi(r_0) P(r_A|r_0; \beta_5, \alpha_5) P(r_B|r_0; \beta_6, \alpha_6) \\ &\times \quad P(Y_1|r_A; \beta_1) P(Y_2|r_A; \beta_2) P(Y_3|r_B; \beta_3) P(Y_4|r_B; \beta_4). \end{aligned} \tag{6}$$

(6) can be further simplified for every specific combination of Y_1, Y_2, Y_3, and Y_4 gene content indices. For instance, the probability of $(Y_1, Y_2, Y_3, Y_4) = (0, 0, 0, 0)$ is

$$\begin{aligned} P(0, 0, 0, 0) \quad &= \quad \sum_{r_0=1}^{\infty} \pi(r_0) \sum_{r_A=0}^{\infty} P(r_A|r_0; \alpha_5, \beta_5) \beta_1^{r_A} \beta_2^{r_A} \sum_{r_B=0}^{\infty} P(r_B|r_0; \alpha_6, \beta_6) \beta_3^{r_B} \beta_4^{r_B} \\ &= \quad \sum_{r_0=1}^{\infty} (1-f)^{r_0} f [\frac{\beta_5 + (1 - \alpha_5 - \beta_5)\beta_1\beta_2}{1 - \alpha_5\beta_1\beta_2}]^{r_0} [\frac{\beta_6 + (1 - \alpha_6 - \beta_6)\beta_3\beta_4}{1 - \alpha_6\beta_3\beta_4}]^{r_0} \\ &= \quad \frac{f\gamma_0}{1 - (1-f)\gamma_0}, \end{aligned}$$

where γ_0 is a function of the parameters $\{\alpha_5, \alpha_6, \beta_1, \cdots, \beta_6\}$, and

$$\gamma_0 = [\frac{\beta_5 + (1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}][\frac{\beta_6 + (1-\alpha_6-\beta_6)\beta_3\beta_4}{1-\alpha_6\beta_3\beta_4}].$$

The second equation is from two probability generating functions for the random variable X_t with each in the form of

$$G_X(s;t) = \sum_{r=0}^{\infty} P(r|r_0;\alpha,\beta)s^r = [\frac{\beta + (1-\alpha-\beta)s}{1-\alpha s}]^{r_0},$$

where α and β are the proliferation and loss parameters, and the value of s is between 0 and 1. The last equation uses the summation of geometric sequences.

The probability mass functions for other situations can be obtained in a similar way, and the results are listed in Table 1. The second column in Table 1 displays the expressions of $\gamma_i, i = 0, 1, \cdots, 14$, used in the probability mass functions. Note that each γ_i is a function of the parameters, $\{\alpha_5, \alpha_6, \beta_1, \cdots, \beta_6\}$.

The gene content configuration of a gene family having been lost in all genomes, i.e. $(Y_1, Y_2, Y_3, Y_4) = (0, 0, 0, 0)$, can not be observed. We standardize the probabilities as

$$q(Y_1, Y_2, Y_3, Y_4) = \frac{P(Y_1, Y_2, Y_3, Y_4)}{1 - P(0, 0, 0, 0)}, \quad (Y_1, Y_2, Y_3, Y_4) \neq (0, 0, 0, 0) \qquad (7)$$

If there are M gene families presented in the four genomes, and we assume independent evolution between lineages, the likelihood function for the parameters, $\{\alpha_5, \alpha_6, \beta_1, \cdots, \beta_6\}$, is

$$\begin{aligned} L(\alpha_1, \alpha_2, \beta_1, \ldots, \beta_6|\text{data}) &= \prod_{m=1}^{M} q(Y_{m1}, Y_{m2}, Y_{m3}, Y_{m4}) \\ &= \prod_{i=1}^{15} q_i^{n_i}, \end{aligned} \qquad (8)$$

where q_i denotes the standardized probability for the i^{th} gene content situation as indicated in the second column of Table 1, and n_i denotes the number of gene families that has pattern i with $\sum_i^{15} n_i = M$. The parameters can then be estimated based on the maximum likelihood approach. Note that the likelihood function (8) does not involve parameters $\alpha_1, \cdots, \alpha_4$. Therefore these parameters are not estimable.

Any four genomes can make 15 rooted trees. The maximum likelihood (ML) for each tree can be found by using, for instance, the Newton-Raphson algorithm. By comparing the 15 calculated MLs, we select a tree with the maximum ML as the reconstructed tree for the four genomes. If there are more than one best trees, we randomly select one among the best ones.

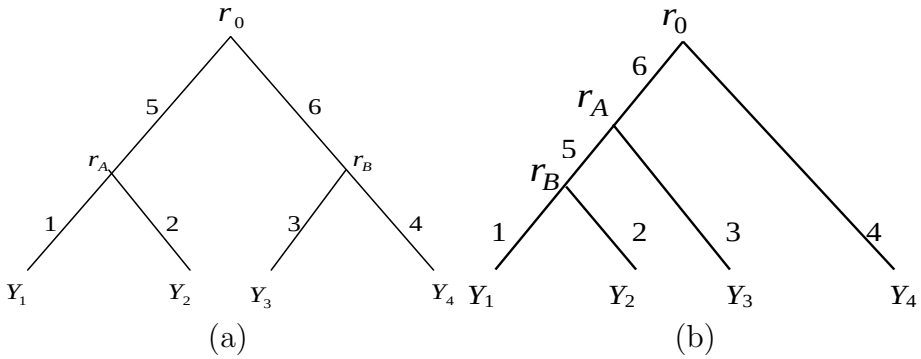

Figure 1: Shapes of the phylogenetic trees of four genomes: (a) symmetric pattern and (b) asymmetric pattern

Note: the numbers, $\{1, 2, \cdots, 6\}$, in the figure are branch indices.

Simulations

We use simulated data sets to demonstrate the developed methodology for phylogeny inference containing four genomes. The tree is simulated based on the shape displayed in Figure 1(a), and assuming that the four external branches in the tree are under the "gene-content clock" situation, i.e. the rates of gene proliferation and loss are the same for all four external branches. Therefore, $\alpha_i's, i = 1, \cdots, 4$, are all the same, and $\beta_i's, i = 1, \cdots, 4$, are all the same as well. Three cases are considered. Case (1) is for total loss, case (2) is for the situation that the gene proliferation rate is higher than the gene loss rate, and case (3) is for short internal branch lengths. For each case, we consider different numbers of gene families at the origin. The values of $\boldsymbol{\mu_t}=\{\mu t_1, \mu t_2, \cdots, \mu t_6\}$ and $\boldsymbol{\lambda t}=\{\lambda t_1, \lambda t_2, \cdots, \lambda t_6\}$ for each case are shown in Table 2, and the correspondent $\alpha_i, \beta_i, i = 1, 2, \cdots, 6$, can be calculated using (2). For each case, we generate 100 trees. Using the maximum likelihood approach, we reconstruct these 100 trees, and record the proportion of trees identified correctly (correctness rate). The correctness rates are displayed in the third column of Table 2. The table shows as long as the internal branch

length is not too short, e.g. as in cases (1) and (2), the correctness rates are normally high (higher than 0.95) even with 50 gene families at the origin. For short internal branch length situation, as shown in Table 2 for case (3), the correctness rate can be significantly improved if we increase the number of gene families.

The results provided here match the results from a gene content approach proposed by Gu and Zhang (2004), and discussed and investigated by Zhang and Gu (2004), in which the trees are constructed through the comparison of pair-wise genome distances calculated based on gene contents. The results also match the results from an extended gene content (absence, single copy, or duplicates of gene families) approach proposed by Gu and Zhang (2004). Zhang and Gu (2004) has a detailed discussion about the comparison between these approaches. Other than these two approaches with which we compare the maximum likelihood approach, we didn't conduct simulation studies for other older approaches. However, the following real data application implies that the likelihood approach is competitive compared to other approaches.

Example

To demonstrate the maximum likelihood approach for four genomes, we apply this approach to infer a phylogenetic tree of 4 complete genomes. We select four genomes from the COG database (http://www.ncbi.nlm.nih.gov/COG/), one Archaea, Archaeoglobus fulgidus (Afu), one Eukaryota, Saccharomyces cerevisiae (Sce), and two Bacteria, Synechocystissp and Helicobacter pylori (Syn and Hpy). Since we already know the relationship among these four genomes, we are able to evaluate reconstructed phylogenetic trees. Totally, there are 2785 gene families considered. Table 3 lists the frequencies for the 15 observable gene content configurations in the order of $\{Afu, Sce, Syn, Hpy\}$. The frequencies are listed in an ascending order. $(0, 1, 1, 0)$ appears least frequently, which implies a large distance between Sce and Syn. $(0, 0, 1, 0)$ is with highest frequency of appearance, and this indicates that Syn may be the closest to the origin among the four genomes.

We maximize the likelihood (8) using the Newton-Raphson algorithm for each possible pattern of the four genomes, and then pick the topology with the maximum ML as the reconstructed tree. The maximum log-likelihoods for the 15 possible patterns are listed in Table 4, and the estimated parameters for the final tree are listed in Table 5. The standard errors are obtained based on the inverse of the Fisher information matrix (Casella and Berger 2001). As indicated by Table 4, the maximum likelihood for the shape in Figure 1(a) with

$\{Y_1, Y_2, Y_3, Y_4\} = \{Afu, Sce, Syn, Hpy\}$ is the largest. This topology is then selected as the phylogenetic tree for the four genomes (Figure 2). Since the additive genome distances are not estimable under our model, in the figure, we use the estimated values of β_i's to roughly reflect the distance between genomes. This overall genome tree supports the concept of universal tree of life (Doolittle 1999), and is similar to the extended gene content tree (Gu and Zhang 2004), the gene content tree (Snel et al. 1999) and the standard 16s RNA tree (Olsen et al. 1994). The reconstructed tree groups the Bacteria (Syn and Hpy) together, and separates Bacteria from the Archaea (Afu) and the Eukaryota (Sce).

As we have seen above, using the maximum likelihood approach can give us the best tree among all possible candidates. Table 4 also indicates another two phylogenetic trees whose maximum likelihoods are very close to the maximum likelihood of the selected tree. The uncertainty of the four genome tree can not be evaluated directly from this approach. To check the uncertainty of the identified phylogenetic tree, we may use the method of bootstrap that is commonly used for other types of genetic data. Each data set used in the bootstrap can be formed by randomly selecting 2785 gene families from the available 2785 ones together with their gene content information for the four genomes.

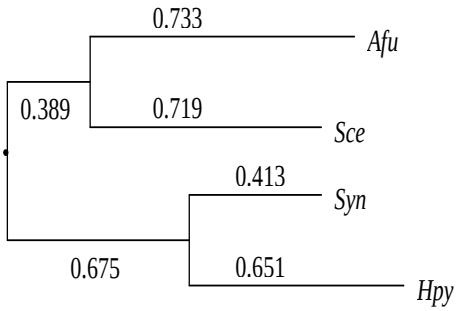

Figure 2: The selected phylogenetic tree for the four genomes using the maximum likelihood approach

Discussion

In this paper, we developed the likelihood function for observing a specific phylogenetic pattern based on four genome situation using gene content in-

formation. Based on the comparison of maximum likelihoods for different phylogenetic trees, we select an optimal tree for the four genomes such that the maximum likelihood of the selected tree is the largest. Simulation results indicate that the proposed method works and can identify trees with a high correctness rate. Through a small example, we demonstrate that the tree of four genomes reconstructed by the proposed approach is acceptable and supports the concept of universal tree of life.

Since the information used for tree-making is only present or absent of a gene family, regardless of the number of member genes, it is believed that this type of approach is robust against gene (genome) duplications, orthologous identification, lateral gene transfer events, and less sensitivity to the algorithm and/or cut-off value used for gene family clustering.

The developed maximum likelihood approach to identifying the phylogenetic tree of four genomes serves as the basis for reconstructing a tree of more than four genomes. The model we set up in this paper for phylogenetic trees is in a simplified version, since the complexity of lateral gene transfer and other issues are overlooked. Nevertheless, the current work can be considered as a start-point toward a rigorous study about complicated issues.

References

[1] Casella, G., Berger, R.L. (2001). Statistical Inference. Duxbury Advanced Series.

[2] Chiang, C. (1968). Introduction to Stochastic Processes in Biostatistics. John Wiley & Sons, Inc.

[3] Clarke G.D.P., Beiko R.G., Ragan, M.A. and Charlebois, R.L. (2002) J. Bacteriol. 184, 2072-2080.

[4] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17: 368-376.

[5] Fitz-Gibbon, S. T., House, C. H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Research: 4218-4222.

[6] Gu, X. (2000). A simple evolutionary model for genome phylogeny inference based on gene content. Comparative genomics (ed. Sankoff and Nadeau) P.515-524. Kluwer Academic Publishers.

[7] Gu, X., Zhang, H. (2004). Genome phylogenetic analysis based on extended gene contents. Mol Biol Evol 21: 1401-1408.

[8] House, C.H. and Fitz-Gibbon, S.T. (2002). Using homolog groups to create a whole-genomic tree of free-living organisms: an update. Journal of Molecular Evolution, 53: 539-547.

[9] Huson, D. H. and Steel M. (2004). Phylogenetic trees based on gene content. Bioinformatics Vol. 20, no. 13: 2044-2049

[10] Huynen, M. A., Snel, B., Bork, P. (1999). Technical comments on Doolittle (1999a), Science 286:1443a.

[11] Korbel, J. O., Snel, B., Huynen, M. A., Bork, P. (2002). SHOT: a web server for the construction of genome phylogenies. Trends Genet. 18:158-62.

[12] Lin, J. and Gerstein, M. (2000). Whole-genome Trees Based on the Occurrence of Folds and Orthologs: Implications for Comparing Genomes on Different Levels. Genome Res. 10: 808-818.

[13] Natale DR, Kidder GM, Westhusin ME, Watson AJ (2000). Assessment of differential display-RT-PCR of mRNA transcript transitions and alpha-amanitin sensitivity during bovine preattachment development. Mol Reprod Dev 55:152-163.

[14] Olsen, G. J., Woese, C. R., and Overbeek, R. (1994). J. Bacteriol. 176, 1-6.

[15] Snel, B., Bork, P., Huynen, M. A. (1999). Genome phylogeny based on gene content. Nat Genet 21:108-110.

[16] Strimmer, K., Haeseler A. (1996). Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies. Mol. Biol. Evol. 13 (7): 964-969

[17] Wolf, Y., I. B. Rogozin, N. V. Grishin and E. V. Koonin (2002). Genome trees and the tree of life, Trends in Genetics, 18: 472-479.

[18] Zhang, H., Gu, X., (2004). Simulation Studies for Genome Phylogeny Inference Based on Gene Contents. Unpublished Technical Report. Department of Mathematics and Statistics, University of West Florida.

Appendix 1

The marginal joint probability of (Y_1, Y_2, Y_3, Y_4) for the shape as indicated in Figure 1(b) is given as,

$$\begin{aligned} P(Y_1, Y_2, Y_3, Y_4) &= \sum_{r_0=1}^{\infty}\sum_{r_A=0}^{\infty}\sum_{r_B=0}^{\infty} \pi(r_0)P(r_A|r_0;\beta_6,\alpha_6)P(r_B|r_A;\beta_5,\alpha_5) \\ &\times \; P(Y_1|r_B;\beta_1)P(Y_2|r_B;\beta_2)P(Y_3|r_A;\beta_3)P(Y_4|r_0;\beta_4). \quad (9) \end{aligned}$$

For $(Y_1, Y_2, Y_3, Y_4) = (0, 0, 0, 0)$, we have

$$\begin{aligned} P(0,0,0,0) &= \sum_{r_0=1}^{\infty}\pi(r_0)\beta_4^{r_0}\sum_{r_A=0}^{\infty}P(r_A|r_0;\alpha_6,\beta_6)\beta_3^{r_A}\sum_{r_B=0}^{\infty}P(r_B|r_A;\alpha_5,\beta_5)\beta_1^{r_B}\beta_2^{r_B} \\ &= \sum_{r_0=1}^{\infty}(1-f)^{r_0}f\beta_4^{r_0}\sum_{r_A=0}^{\infty}P(r_A|r_0;\alpha_6,\beta_6)\beta_3^{r_A}[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}]^{r_A} \\ &= \sum_{r_0=1}^{\infty}(1-f)^{r_0}f\beta_4^{r_0}[\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3\delta_0}{1-\alpha_6\beta_3\delta_0}]^{r_0} \\ &= \frac{f\gamma_0}{1-(1-f)\gamma_0}, \end{aligned}$$

where $\delta_0 = \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}$, and γ_0 is a function of the parameters $\{\alpha_5, \alpha_6, \beta_1, \cdots, \beta_6\}$ such that

$$\gamma_0 = \beta_4 \times [\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}}{1-\alpha_6\beta_3\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}}]$$

Using (9), the marginal joint probabilities for other situations can be derived in the same way. These probabilities have the same format as those for the shape indicated in Figure 1(a) but with different $\gamma_i, i = 0, \cdots, 14$. The calculation formulas for γ_i's are listed in Table 6.

Similarly, to overcome unobservable double-losses, we standardize $P(Y_1, Y_2, Y_3, Y_4)$. The likelihood function for the parameters, $\{\alpha_5, \alpha_6, \beta_1, \cdots, \beta_6\}$, with M gene families in the four genomes is in the same format as that given by (8), which

is

$$\begin{aligned} L(\beta_1, \ldots, \beta_6|\text{data}) &= \prod_{m=1}^{M} q(Y_{m1}, Y_{m2}, Y_{m3}, Y_{m4}) \\ &= \prod_{i=1}^{15} q_i^{n_i}, \end{aligned} \tag{10}$$

where q_i is the standardized probability for the i_{th} gene content situation as indicated in the first column of Table 6, and n_i denotes the number of gene families that has pattern i with $\sum_i^{15} n_i = M$.

Table 1: The probability mass function (pmf) of $\mathbf{Y} = \{Y_1, Y_2, Y_3, Y_4\}$

i	$\mathbf{Y}$	γ_i	pmf
0	(0,0,0,0)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}][\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3\beta_4}{1-\alpha_6\beta_3\beta_4}]$	$\frac{f\gamma_0}{1-(1-f)\gamma_0}$
1	(0,0,0,1)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}][\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3)}{1-\alpha_6\beta_3}]$	$\frac{f\gamma_1}{1-(1-f)\gamma_1} - \frac{f\gamma_0}{1-(1-f)\gamma_0}$
2	(0,0,1,0)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}][\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_4}{1-\alpha_6\beta_4}]$	$\frac{f\gamma_2}{1-(1-f)\gamma_2} - \frac{f\gamma_0}{1-(1-f)\gamma_0}$
3	(0,1,0,0)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}][\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3\beta_4}{1-\alpha_6\beta_3\beta_4}]$	$\frac{f\gamma_3}{1-(1-f)\gamma_3} - \frac{f\gamma_0}{1-(1-f)\gamma_0}$
4	(1,0,0,0)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}][\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3\beta_4}{1-\alpha_6\beta_3\beta_4}]$	$\frac{f\gamma_4}{1-(1-f)\gamma_4} - \frac{f\gamma_0}{1-(1-f)\gamma_0}$
5	(0,1,1,0)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}][\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_4}{1-\alpha_6\beta_4}]$	$\frac{f\gamma_5}{1-(1-f)\gamma_5} - \frac{f\gamma_3}{1-(1-f)\gamma_3} - \frac{f\gamma_2}{1-(1-f)\gamma_2} + \frac{f\gamma_0}{1-(1-f)\gamma_0}$
6	(0,0,1,1)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}]$	$\frac{f\gamma_6}{1-(1-f)\gamma_6} - \frac{f\gamma_2}{1-(1-f)\gamma_2} - \frac{f\gamma_1}{1-(1-f)\gamma_1} + \frac{f\gamma_0}{1-(1-f)\gamma_0}$
7	(0,1,0,1)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}][\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3}{1-\alpha_6\beta_3}]$	$\frac{f\gamma_7}{1-(1-f)\gamma_7} - \frac{f\gamma_3}{1-(1-f)\gamma_3} - \frac{f\gamma_1}{1-(1-f)\gamma_1} + \frac{f\gamma_0}{1-(1-f)\gamma_0}$
8	(1,0,0,1)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}][\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3}{1-\alpha_6\beta_3}]$	$\frac{f\gamma_8}{1-(1-f)\gamma_8} - \frac{f\gamma_4}{1-(1-f)\gamma_4} - \frac{f\gamma_1}{1-(1-f)\gamma_1} + \frac{f\gamma_0}{1-(1-f)\gamma_0}$
9	(1,1,0,0)	$[\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3\beta_4}{1-\alpha_6\beta_3\beta_4}]$	$\frac{f\gamma_9}{1-(1-f)\gamma_9} - \frac{f\gamma_3}{1-(1-f)\gamma_3} - \frac{f\gamma_4}{1-(1-f)\gamma_4} + \frac{f\gamma_0}{1-(1-f)\gamma_0}$
10	(1,0,1,0)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}][\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_4}{1-\alpha_6\beta_4}]$	$\frac{f\gamma_{10}}{1-(1-f)\gamma_{10}} - \frac{f\gamma_4}{1-(1-f)\gamma_4} - \frac{f\gamma_2}{1-(1-f)\gamma_2} + \frac{f\gamma_0}{1-(1-f)\gamma_0}$
11	(0,1,1,1)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}]$	$\frac{f\gamma_{11}}{1-(1-f)\gamma_{11}} - \frac{f\gamma_7}{1-(1-f)\gamma_7} - \frac{f\gamma_5}{1-(1-f)\gamma_5} + \frac{f\gamma_3}{1-(1-f)\gamma_3} - P(0,0,1,1)$
12	(1,1,1,0)	$[\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_4}{1-\alpha_6\beta_4}]$	$\frac{f\gamma_{12}}{1-(1-f)\gamma_{12}} - \frac{f\gamma_5}{1-(1-f)\gamma_5} - \frac{f\gamma_{10}}{1-(1-f)\gamma_{10}} + \frac{f\gamma_2}{1-(1-f)\gamma_2} - P(1,1,0,0)$
13	(1,1,0,1)	$[\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3}{1-\alpha_6\beta_3}]$	$\frac{f\gamma_{13}}{1-(1-f)\gamma_{13}} - \frac{f\gamma_8}{1-(1-f)\gamma_8} - \frac{f\gamma_7}{1-(1-f)\gamma_7} + \frac{f\gamma_1}{1-(1-f)\gamma_1} - P(1,1,0,0)$
14	(1,0,1,1)	$[\frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}]$	$\frac{f\gamma_{14}}{1-(1-f)\gamma_{14}} - \frac{f\gamma_{10}}{1-(1-f)\gamma_{10}} - \frac{f\gamma_8}{1-(1-f)\gamma_8} + \frac{f\gamma_4}{1-(1-f)\gamma_4} - P(0,0,1,1)$
15	(1,1,1,1)	—	$1 - (\frac{f\gamma_{13}}{1-(1-f)\gamma_{13}} + \frac{f\gamma_{12}}{1-(1-f)\gamma_{12}} - \frac{f\gamma_9}{1-(1-f)\gamma_9}) - P(0,1,1,1) - P(1,0,1,1) - P(0,0,1,1)$

Table 2: The correctness rates for different types of trees

Cases	# gene families	Correctness rate
(1)	50	0.95
	100	0.97
	200	1
(2)	50	0.96
	100	0.99
	200	1
(3)	50	0.34
	100	0.47
	200	0.53
	500	0.69

Parameters for each case:
case (1): $\boldsymbol{\mu t}$=$\{\mathbf{0.5}\}$, $\boldsymbol{\lambda t}$=$\{\mathbf{0}\}$
case (2): $\boldsymbol{\mu t}$=$\{\mathbf{0.1}\}$, $\boldsymbol{\lambda t}$=$\{\mathbf{0.12}\}$
case (3): $\boldsymbol{\mu t}$=$\{0.1/16, 0.1/16, 0.1, 0.1, 0.1, 0.1\}$
$\boldsymbol{\lambda t}$=$\{0.12/16, 0.12/16, 0.12, 0.12, 0.12, 0.12\}$

Table 3: Frequencies for the 0/1 pattern of $\{Afu, Sce, Syn, Hpy\}$

$\{Y_1, Y_2, Y_3, Y_4\}$	Frequency
$(0, 1, 1, 0)$	21
$(1, 1, 1, 0)$	22
$(1, 1, 0, 0)$	39
$(1, 1, 0, 1)$	90
$(1, 0, 1, 1)$	113
$(0, 0, 1, 1)$	123
$(0, 1, 1, 1)$	142
...	...
$(0, 0, 0, 1)$	388
$(0, 0, 1, 0)$	481

Table 4: Maximum likelihood for each possible topology of the four genomes

Possible patterns	Max. Log-likelihood
Topology in Figure 1(a)	
(Y_1, Y_2, Y_3, Y_4)	-6865.9
(Y_1, Y_3, Y_2, Y_4)	-7071.8
(Y_1, Y_4, Y_2, Y_3)	-7055.8
Topology in Figure 1(b)	
(Y_1, Y_2, Y_3, Y_4)	-7061.3
(Y_1, Y_3, Y_2, Y_4)	-7080.4
(Y_2, Y_3, Y_1, Y_4)	-7038.4
(Y_1, Y_2, Y_4, Y_3)	-7075.4
(Y_1, Y_4, Y_2, Y_3)	-7081.9
(Y_2, Y_4, Y_1, Y_3)	-7086.6
(Y_1, Y_4, Y_3, Y_2)	-7001.8
(Y_1, Y_3, Y_4, Y_2)	-7002.4
(Y_4, Y_3, Y_1, Y_2)	-6866.8
(Y_4, Y_2, Y_3, Y_1)	-7017.9
(Y_4, Y_3, Y_2, Y_1)	-6872.0
(Y_2, Y_3, Y_4, Y_1)	-7016.8

Note: The four genomes $(Y_1, Y_2, Y_3, Y_4) = (Afu,\ Sce,\ Syn,\ Hpy)$

Table 5: Estimated parameters together with their standard errors

	α_5	α_6	β_1	β_2	β_3	β_4	β_5	β_6
Estimate	0.000001	0.043	0.733	0.719	0.413	0.651	0.389	0.675
Std. Err.	0	0.0347	0.0076	0.0079	0.0137	0.0105	0.0155	0.00862

Table 6: The expressions of $\gamma_i, i = 0, 1, \cdots, 14$

$\mathbf{Y}$	i	γ_i
(0,0,0,0)	0	$\beta_4 \times [\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}}{1-\alpha_6\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}}]$
(0,0,0,1)	1	$\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}}{1-\alpha_6\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}}$
(0,0,1,0)	2	$\beta_4 \times [\frac{\beta_6+(1-\alpha_6-\beta_6) \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}}{1-\alpha_6 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}}]$
(0,1,0,0)	3	$\beta_4 \times [\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}}{1-\alpha_6\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}}]$
(1,0,0,0)	4	$\beta_4 \times [\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}}{1-\alpha_6\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}}]$
(0,1,1,0)	5	$\beta_4 \times [\frac{\beta_6+(1-\alpha_6-\beta_6) \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}}{1-\alpha_6 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}}]$
(0,0,1,1)	6	$\frac{\beta_6+(1-\alpha_6-\beta_6) \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}}{1-\alpha_6 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1\beta_2}{1-\alpha_5\beta_1\beta_2}}$
(0,1,0,1)	7	$\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}}{1-\alpha_6\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}}$
(1,0,0,1)	8	$\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}}{1-\alpha_6\beta_3 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}}$
(1,1,0,0)	9	$\beta_4 \times [\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3}{1-\alpha_6\beta_3}]$
(1,0,1,0)	10	$\beta_4 \times [\frac{\beta_6+(1-\alpha_6-\beta_6) \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}}{1-\alpha_6 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}}]$
(0,1,1,1)	11	$\frac{\beta_6+(1-\alpha_6-\beta_6) \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}}{1-\alpha_6 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_1}{1-\alpha_5\beta_1}}$
(1,1,1,0)	12	β_4
(1,1,0,1)	13	$\frac{\beta_6+(1-\alpha_6-\beta_6)\beta_3}{1-\alpha_6\beta_3}$
(1,0,1,1)	14	$\frac{\beta_6+(1-\alpha_6-\beta_6) \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}}{1-\alpha_6 \frac{\beta_5+(1-\alpha_5-\beta_5)\beta_2}{1-\alpha_5\beta_2}}$