Genome Phylogenetic Analysis Based on Extended Gene Contents

Xun Gu,† and Hongmei Zhang†‡*

*Department of Genetics, Development, and Cell Biology, and †Center for Bioinformatics and Biological Statistics, Iowa State University; ‡Department of Mathematics and Statistics, University of West Florida

With the rapid growth of entire genome data, whole-genome approaches such as gene content become popular for genome phylogeny inference, including the tree of life. However, the underlying model for genome evolution is unclear, and the proposed (ad hoc) genome distance measure may violate the additivity. In this article, we formulate a stochastic framework for genome evolution, which provides a basis for defining an additive genome distance. However, we show that it is difficult to utilize the typical gene content data—i.e., the presence or absence of gene families across genomes—to estimate the genome distance. We solve this problem by introducing the concept of extended gene content; that is, the status of a gene family in a given genome could be absence, presence as single copy, or presence as duplicates, any of which can be used to estimate the genome distance and phylogenetic inference. Computer simulation shows that the new tree-making method is efficient, consistent, and fairly robust. The example of 35 microbial complete genomes demonstrates that it is useful not only to study the universal tree of life but also to explore the evolutionary pattern of genomes.

Introduction

Since the concept of the tree of life was proposed (Woese 1987), it was thought that more sequences of orthologous genes could improve the depth and resolution of our knowledge of life's history. This view has been challenged since the publication of the first microbial genome sequence, *Haemophilus influenzae*. To date the roster of complete genomes is close to 100 (for an overview, see http://www.tigr.org). In spite of more than 10 prokaryotic phyla plus a few eukaryotes represented, we are actually facing more difficulties in having a meaningful interpretation of the tree of life. Because phylogenetic analysis based on a single gene (family) has produced many conflicted gene trees, the long-term controversy over "vertical" (tree-like) evolution versus lateral (horizontal) gene transfer has become more heated rather than resolved in the genome era (Golding and Gupta 1995; Doolittle and Logsdon 1998; Jain, Rivera, and Lake 1999; Doolittle 1999*a*, 1999*b*; Nelson et al. 1999; Tekaia, Lazcano, and Dujon 1999; Huynen and Snel 2000; Wolf et al. 2002; Daubin, Moran, and Ochman 2003).

Because phylogenetic trees of individual genes are inconsistent, the whole-genome analysis—e.g., the gene content (the presence/absence of gene families over genomes)—is becoming an attractive approach to extracting the bulk phylogenetic signals. For instance, several authors (Snel, Bork, and Huynen 1999; Huynen, Snel, and Bork 1999; Lin and Gerstein 2000; Korbel et al. 2002) estimated the fraction of shared genes for genome pairs and transformed that fraction 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).

Interestingly, these genome-level studies show a general similarity between the gene-content tree and the classical rRNA tree, implying that the vertical (tree-like) evolutionary history of an organism could be maintained at the genome level, which is not seriously affected by the lateral gene transfer. However, Doolittle (1999*b*) raised a fundamental question about whether a genome tree based on gene content alone, and not the evolutionary relationship, is the best phenotypic measure. In fact, any inferred topology (including molecular phylogeny) could be potentially misleading. For instance, the high variation of the GC% in bacterial genomes results in high variation of amino acid compositions (Gu 2001) that may complicate the phylogenetic inference based on protein sequences. An inferred topology turns out to be an estimate of the phylogenetic relationship only when the assumptions have been carefully examined. A common problem shared by these genome approaches is the lack of a clear-cut evolutionary model. Consequently, these studies at best lead to a much weaker statement: that the genome tree might be interpreted as only a prevailing trend in the evolution of genome-scale gene sets rather than as a dominant picture of evolution (Wolf et al. 2002).

We have recognized the important role of modeling for phylogenomic analysis in justifying whether the inferred tree indeed represents the genome phylogeny. Because the likelihood framework for phylogenetic gene-content analysis (Gu 2000) may require a huge amount of computational time, the genome distance approach is demanding in practice. In this article, we first show that the gene-content distance is generally not additive, so its application for phylogenomic analysis could be misleading. We then tackle this problem by extending the concept of gene content into a more general framework such that the additive genome distance can be estimated. The efficiency of genome phylogenetic reconstruction is examined by extensive computer simulations. Finally, we apply the newly developed method to study the universal tree of life.

The Stochastic Model

The Joint Size Distribution of the Gene Family in Multiple Genomes

The whole-genome comparison has revealed a high variation in the size of gene families among complete genomes, because a gene family can be generated,

Key words: Gene content; additive genome distance; phylogenetic inference; comparative genomics.

E-mail: xgu@iastate.edu

Mol. Biol. Evol. 21(7):1401–1408. 2004
doi:10.1093/molbev/msh138
Advance Access publication April 14, 2004

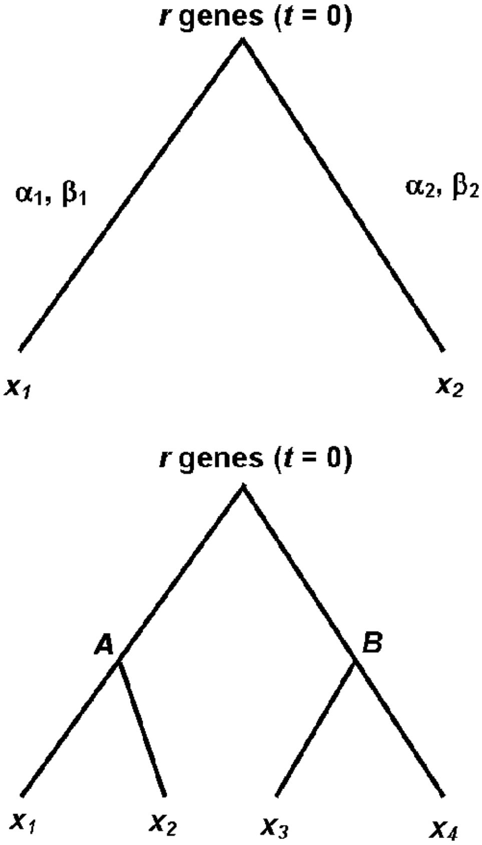

FIG. 1.—Schematic genome evolution for two genomes and four genomes, respectively. The gene family has r member genes in the root. After t evolutionary time units, the size of the gene family is x_1 and x_2 in genomes 1 and 2, respectively. For four genomes, the size of the gene family is x_i ($i = 1, \ldots, 4$).

expanded, reduced, or lost during the course of genome evolution. Therefore, the joint size distribution of the gene family among genomes is useful for phylogenomic analysis.

Nei et al. (1997) proposed a birth-death hypothesis for the evolution of young duplicate genes. Here we develop a general stochastic model, considering two major evolutionary processes that influence the size of a gene family: gene loss (nonfunctionization or deletion) and gene proliferation (duplication). Let μ be the evolutionary rate of gene loss and λ be the evolutionary rate of gene proliferation. If each gene is subject to the same chance of being lost or duplicated, for a gene family with r member genes at $t = 0$, the number of member genes after t time units, denoted by X_t, follows the following distribution

$$\begin{aligned} P(X_t = k \mid X_0 = r) &= \sum_{j=0}^{\min[r,k]} \binom{r}{j}\binom{r+k-j-1}{r-1} \\ &\quad \times \beta^{r-j}\alpha^{k-j}(1-\alpha-\beta)^j, \quad k \geq 1 \\ P(X_t = 0 \mid X_0 = r) &= \beta^r, \end{aligned} \tag{1}$$

where the proliferation parameter α and the loss parameter β are given by

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

respectively. Equation (2) implies $\alpha / \beta = \lambda / \mu$, which is called the P / L ratio. The size of the gene family under the birth-death model is expected to be $X_0e^{(\lambda-\mu)t}$, $\alpha > \beta$ (or $P / L > 1$), which indicates, on average, an increase of gene family size during evolution and vice versa.

Consider two genomes that diverged t time units ago (fig. 1). For a given gene family, we assume that there are r member genes at $t = 0$ (in the common ancestor). Let X_i, $i = 1,2$, denote the number of genes after t time units for genome i. Under the assumption of independent evolution between lineages, the (conditional) joint probability is given by $P(X_1, X_2 \mid X_0 = r) = P(X_1 \mid X_0 = r) \times P(X_2 \mid X_0 = r)$. Because the size of a gene family in the ancestral genome is unknown, a (prior) distribution for $X_0 = r$ is assumed, denoted by $\pi(r)$. Thus, the joint probability of X_1 and X_2 is given by

$$\begin{aligned} P(X_1, X_2) &= \sum_{r=1}^{\infty} \pi(r)P(X_1, X_2 \mid X_0 = r) \\ &= \sum_{r=1}^{\infty} \pi(r)P(X_1 \mid r)P(X_2 \mid r), \end{aligned} \tag{3}$$

where $P(X_i \mid r)$ is short for $P(X_i \mid X_0 = r)$ defined by equation (1).

For the general n-genomes, let X_i represent the size of a gene family in the ith genome, $i = 1, \ldots, n$. The joint size distribution of the gene family $\mathbf{X} = (X_1, \ldots, X_n)$ can be derived according to the Markov chain model, similar to DNA sequence evolution (Felsenstein 1981). For example, for four genomes (fig. 1), it is given by

$$\begin{aligned} P(\mathbf{X}) = &\sum_{r_0}\sum_{r_A}\sum_{r_B} \pi(r_0)P(r_A \mid r_0; \alpha_5, \beta_5)P(r_B \mid r_0; \alpha_6, \beta_6) \\ &\times P(X_1 \mid r_A; \alpha_1, \beta_1)P(X_2 \mid r_A; \alpha_2, \beta_2) \\ &\times P(X_3 \mid r_B; \alpha_3, \beta_3)P(X_4 \mid r_B; \alpha_4, \beta_4), \end{aligned} \tag{4}$$

where $P(. \mid .; \alpha_i, \beta_i)$ is the transition probability for branch i, defined by equation (1).

Two-Genome Model and Expression Distance

The Additive Genome Distance Measures

Given the joint-size distribution, say, equation (4) for four genomes, maximum likelihood phylogeny can be implemented. Unfortunately, the complexity of transition probability (eq. 1) makes it almost intractable for the genome-level analysis. Thus, the distance method becomes highly desirable, but first one should define an additive genome distance measure. With some algebras from equation (2), two quantities, the proliferation measure d_λ and the loss measure d_μ, are given by

$$d_\lambda = \frac{\alpha}{\beta-\alpha}\ln\frac{1-\alpha}{1-\beta} = \lambda t$$

and

$$d_\mu = \frac{\beta}{\beta-\alpha}\ln\frac{1-\alpha}{1-\beta} = \mu t, \tag{5}$$

respectively. For two genomes (fig. 1), let λ_i, μ_i, α_i, β_i, d_{λ_i} and d_{μ_i} be the corresponding parameters in each lineage, $i = 1, 2$; see equations (2) and (5). Then we define the proliferation genome distance between two genomes (the P distance, for short) as $G_P = d_{\lambda_1} + d_{\lambda_2} = (\lambda_1 + \lambda_2)t$; from equation (5), it is given by

$$G_P = \sum_{i=1,2} \frac{\alpha_i}{\beta_i - \alpha_i} \ln \frac{1-\alpha_i}{1-\beta_i}. \qquad (6)$$

In the same manner, the loss genome distance (L distance, for short) between two genomes is defined as $G_L = d_{\mu_1} + d_{\mu_2} = (\mu_1 + \mu_2)t$, given by

$$G_L = \sum_{i=1,2} \frac{\beta_i}{\beta_i - \alpha_i} \ln \frac{1-\alpha_i}{1-\beta_i}, \qquad (7)$$

and the general genome distance measure is defined as $G = G_P + G_L$, i.e.,

$$G = \sum_{i=1,2} \frac{\alpha_i + \beta_i}{\beta_i - \alpha_i} \ln \frac{1-\alpha_i}{1-\beta_i}. \qquad (8)$$

Apparently, these genome distance measures are additive, and $G_P/G_P = P/L$ ratio. Equations (6)–(8) provide the relationship between genome distances and parameters in the probabilistic model (eqs. 1–3). To estimate the genome distance, we shall develop a computationally efficient method for estimating the parameters (α_i and β_i).

Gene Content: It's Not Sufficient

The concept of *gene content* was introduced by several authors for studying the universal genome tree (e.g., Snel, Bork, and Huynen 1999; Tekaia, Lazcano, and Dujon 1999). For two genomes $i = 1, 2$, let Y_i be the gene-content index of a gene family: $Y_i = 1$ indicates at least one member gene found in the ith genome; otherwise $Y_i = 0$. Therefore, gene-content pattern is the most degenerated size distribution of the gene family. In the following discussion we will show that it becomes insufficient for estimating the genome distance.

From equation (3) one can show that the joint probability of Y_1 and Y_2 is given by

$$P(Y_1, Y_2) = \sum_{r=1}^{\infty} \pi(r) P(Y_1 \mid r) P(Y_2 \mid r). \qquad (9)$$

Because $P(Y_i = 0 \mid r) = \beta_i^r$, and $P(Y_i = 1 \mid r) = 1 - \beta_i^r$, $i = 1, 2$, the analytical form of $P(Y_1, Y_2)$ can be obtained if a geometric prior is assumed, i.e., $\pi(r) = (1-f)^{r-1} f$. For simplicity, let $P(i, j) = P(Y_1 = i, Y_2 = j)$. Then, putting $\pi(r)$ into equation (9), we have

$$\begin{aligned} P(1,1) &= 1 - Q(\beta_1) - Q(\beta_2) + Q(\beta_1\beta_2) \\ P(1,0) &= Q(\beta_2) - Q(\beta_1\beta_2) \\ P(0,1) &= Q(\beta_1) - Q(\beta_1\beta_2) \\ P(0,0) &= Q(\beta_1\beta_2), \end{aligned} \qquad (10)$$

where the function $Q(\beta)$ ($\beta = \beta_1$, β_2 or $\beta_1\beta_2$) is defined as

$$Q(\beta) = \sum_{r=1}^{\infty} \pi(r) \beta^r = \frac{\beta f}{1 - (1-f)\beta}. \qquad (11)$$

Because equation (10) relies only on the loss parameters β_1 and β_2, we cannot estimate the proliferation parameters (α_1 and α_2). In other words, the additive genome distances defined by equations (6)–(8) in general cannot be estimated by the gene-content approach.

Extended Gene Content

We have found a plausible solution by further dividing the non-zero (member genes) case into two states: single-copy (one-member) genes or duplicates (more than one member genes). This extended gene-content analysis considers three possible states: no member gene ($Z = 0$), single-copy gene ($Z = 1$), and duplicate genes ($Z = 2$). According to equation (1), their probabilities are $P(Z = 0 \mid X_0 = r) = P(X_t = 0 \mid X_0 = r)$, $P(Z = 1 \mid X_0 = r) = P(X_t = 1 \mid X_0 = r)$ and $P(Z = 2 \mid X_0 = r) = \sum_{k \geq 2} P(X_t = k \mid X_0 = r)$, as given by

$$\begin{aligned} &P(Z = 0 \mid X_0 = r) = \beta^r \\ &P(Z = 1 \mid X_0 = r) = r\beta^{r-1}(1-\beta)(1-\alpha) \\ &P(Z = 2 \mid X_0 = r) = 1 - \beta^r - r\beta^{r-1}(1-\beta)(1-\alpha), \end{aligned} \qquad (12)$$

respectively.

The Joint Distribution for Two Genomes

Consider two genomes that diverged t time units ago (fig. 1). Let $Z_i = 0$, 1, or 2 be the extended gene-content index for a gene family in the ith genome, $i = 1, 2$. Similar to equation (3) and equation (9), the joint distribution of Z_1 and Z_2 is given by

$$\begin{aligned} P(Z_1, Z_2) &= \sum_{r=1}^{\infty} \pi(r) P(Z_1, Z_2 \mid X_0 = r) \\ &= \sum_{r=1}^{\infty} \pi(r) P(Z_1 \mid r) P(Z_2 \mid r), \end{aligned} \qquad (13)$$

where $P(Z_i \mid r) = P(Z_i \mid X_0 = r)$. Given the geometric distribution for $\pi(r) = f(1-f)^{r-1}$, we obtain the analytical forms of equation (13) as follows

$$\begin{aligned} P(0,0) &= Q(\beta_1\beta_2) \\ P(0,1) &= \beta_1\omega_2 R(\beta_1\beta_2) \\ P(0,2) &= Q(\beta_1) - Q(\beta_1\beta_2) - \beta_1\omega_2 R(\beta_1\beta_2) \\ P(1,0) &= \beta_2\omega_1 R(\beta_1\beta_2) \\ P(1,1) &= \omega_1\omega_2 S(\beta_1\beta_2) \\ P(1,2) &= \omega_1[R(\beta_1) - \beta_2 R(\beta_1\beta_2)] - \omega_1\omega_2 S(\beta_1\beta_2) \\ P(2,0) &= Q(\beta_2) - Q(\beta_1\beta_2) - \beta_2\omega_1 R(\beta_1\beta_2) \\ P(2,1) &= \omega_2[R(\beta_2) - \beta_1 R(\beta_1\beta_2)] - \omega_1\omega_2 S(\beta_1\beta_2) \\ P(2,2) &= 1 - Q(\beta_1) - Q(\beta_2) + Q(\beta_1\beta_2) \\ &\quad - \omega_1[R(\beta_1) - \beta_2 R(\beta_1\beta_2)] \\ &\quad - \omega_2[R(\beta_2) - \beta_1 R(\beta_1\beta_2)] + \omega_1\omega_2 S(\beta_1\beta_2), \end{aligned} \qquad (14)$$

where $\omega_1 = (1 - \beta_1)(1 - \alpha_1)$ and $\omega_2 = (1 - \beta_2)(1 - \alpha_2)$; the function $Q(\beta)$ is given by equation (11), the function $R(\beta) = \sum_{r=1}^{\infty} \pi(r) r \beta^{r-1}$ is given by

$$R(\beta) = \frac{f}{1-(1-f)\beta} + \frac{f(1-f)\beta}{[1-(1-f)\beta]^2}, \qquad (15)$$

and the function $S(\beta) = \sum_{r=1}^{\infty} \pi(r) r^2 \beta^{r-1}$ is given by

$$S(\beta) = \frac{f}{1-(1-f)\beta} + \frac{3f(1-f)\beta}{[1-(1-f)\beta]^2} + \frac{2f(1-f)^2\beta^2}{[1-(1-f)\beta]^3}. \qquad (16)$$

Here $\beta = \beta_1$, β_2 or $\beta_1\beta_2$.

Parameter Estimation

When the extended gene-content data matrix for any two genomes 1 and 2 is given, we develop a maximum likelihood (ML)–based approach to estimating the genome distances. Usually the prior parameter f can be estimated from the observed size frequencies of gene families. Because the pattern of double loss (i.e., $Z_1 = 0$ and $Z_2 = 0$) is not observable, one may use the following modified joint probability,

$$q(Z_1, Z_2) = \frac{P(Z_1, Z_2)}{1 - P(0,0)} = \frac{P(Z_1, Z_2)}{1 - Q(\beta_1\beta_2)}, \qquad (17)$$

for Z_1, $Z_2 = 0$, 1 or 2, except $Z_1 = Z_2 = 0$. Let n_{ij} be the number of gene families with the pattern $Z_1 = i$ and $Z_2 = j$, where $i, j = 0$, 1, 2 except $i = j = 0$. Then, the likelihood for the two genomes can be written as

$$L(\alpha_1, \alpha_2, \beta_1, \beta_2 \mid \text{data}) = \prod_{i,j} q(i,j)^{n_{ij}}. \qquad (18)$$

We use the Newton-Raphoson numerical iteration to obtain the ML estimates of α_1, α_2, β_1, and β_2. Their sampling variance-covariance matrix is approximately computed by the inverse of Fisher's information matrix. When these parameters (α_1, α_2, β_1, β_2) are estimated, the computation of genome distances by equations (6)–(8) are straightforward, and the sampling variance of a genome distance can be obtained by the delta method.

Computer Simulations

We have conducted extensive computer simulations to examine the performance of phylogenetic reconstruction using the extended gene-content data. The computer program is encoded using the language C++. The number of replications in each simulation study is set at 2,000. Because of space limitations, we will discuss our main results briefly.

Estimation of Genome Distance Is Asymptotically Unbiased

We first simulate the stochastic process according to the two-genome evolution scenario (fig. 1), when the evolutionary parameters ($\lambda_i t$ and $\mu_i t$, $i = 1, 2$) are given. For each gene family, the number of genes on the root, r, is generated from a geometric distribution with the parameter $f = 0.5$. In each replicate, we implement the ML algorithm to estimate the proliferation parameter α_i and the loss parameter β_i ($i = 1, 2$), and we then compute the genome distances according to equations (6)–(8). The mean and variance for each estimate are used for examining the statistical properties.

We have studied four typical cases: the gene-loss model ($\lambda = 0$), the growth model ($\lambda > \mu$), the equal model ($\lambda = \mu$), and the reduction model ($\lambda < \mu$). The number of gene families (N) is set at N = 200, 500, and 1,000, respectively. We have examined a variety of combinations from these models in two lineages and have found that the estimates of these parameters and genome distances are asymptotically biased, which is virtually trivial when $N >$ 500. The sampling variances of genome distances decrease with the increase in the number of gene families, and the variances are usually acceptable if $N >$ 500.

Genome Tree Inference Is Efficient and Consistent

We have examined the tree-making performance of the extended gene-content approach, using a typical four-genome scenario (fig. 2). After the extended gene-content matrix of four genomes is simulated, we estimate the genome distance matrix and then infer the tree with the Neighbor-Joining (NJ) algorithm. The efficiency of phylogenetic inference is then measured by the percentage of correct topology inference over 1,000 replicates. After having examined many combinations, we concluded that our method is efficient; that is, except in some extreme cases, the correct percentage is satisfactory (>70%) when $N >$ 500. Our method is also consistent; that is, the correct percentage tends to be 100% when $N \to \infty$.

Table 1 shows the correct percentage of tree-making when the true tree has four equal external branch lengths (fig. 2*A*). When the internal branch length (c) is short, the genome tree inference can be significantly improved as N becomes larger. To examine the tree-making consistency, we consider two typical patterns when the external branches are highly unequal (fig. 2B and fig. 2C). As shown in table 2, the performance is poor when N is small and the internal branch length is short. Nevertheless, even in the very extreme case, the correct percentage of tree-making is close to 100% for sufficiently large number of gene families.

We have also investigated the effect of the prior distribution. We use several alternative distributions in our simulation model that have a longer tail than the geometric distribution. For instance, $\pi(r) = C(1-f)^{\sqrt{r-1}} f$, or $\pi(r) = Cr^{-\gamma}$ (C is the normalizing constant). After we examined many cases, we found that the performance of tree-making is very robust against the choice of a specific form of $\pi(r)$ (not shown).

Example: The Universal Genome Tree of Life

To compare it with previous genome phylogenetic inferences using gene-content data, we applied the newly developed extended gene-content method to infer the universal genome tree of 35 complete genomes, similar to

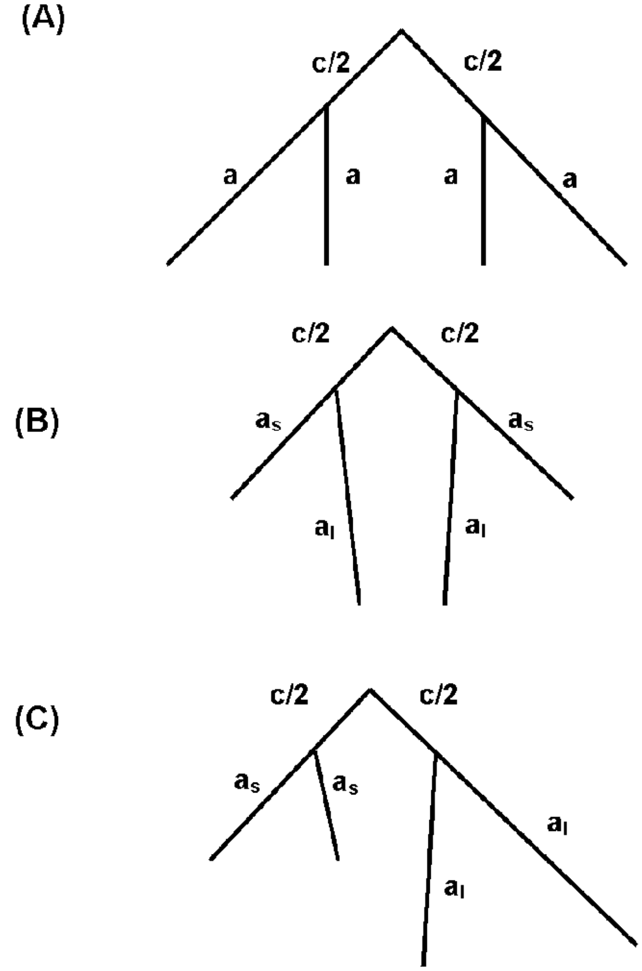

FIG. 2.—The genome tree used for a computer simulation study. *A*. Equal external branch lengths. *B*. Unequal external branches (Felsenstein's zone). *C*. Unequal external branches (non-Felsenstein's zone).

Table 1
Correct Percentage (%) of Tree Making: Equal External Branch Lengths (see fig. 2A)

	The *c*/*a* Ratio				
N	1	1/2	1/4	1/8	1/16
(1) $a = 0.5$, $P/L = 0$					
100	100	95	78	55	50
500	100	100	98	85	59
2,000	100	100	100	99	70
(2) $a = 0.75$, $P/L = 0.5$					
100	100	96	82	57	54
500	100	100	100	95	66
2,000	100	100	100	100	78
(3) $a = 1.0$, $P/L = 1$					
100	100	100	89	63	44
500	100	100	100	88	67
2,000	100	100	100	98	82
(4) $a = 0.75$, $P/L = 2$					
100	100	99	86	64	46
500	100	100	100	91	59
2,000	100	100	100	100	73

Wolf et al. (2002). The extended gene-content data were obtained from the COG database (http://www.ncbi.nlm.nih.gov/COG/). Then, the pairwise genome distance (*G*) was estimated according to equation (8). We also estimated the proliferation (*P*) and the loss (*L*) genome distances, respectively (data not shown).

We used the NJ method (Saitou and Nei 1987) to infer the genome phylogeny. The overall genome tree based on extended gene content (Fig. 3) supports the concept of a universal tree, similar to previous gene-content trees (Snel, Bork, and Huynen 1999; Wolf et al. 2002) and the standard 16s RNA tree (Olsen, Woese, and Overbeek 1994). That is, two major lineages of cellular life, the Archaea and the Bacteria, are monophyletic from the third lineage (Eukarya, represented by the yeast genome), supported by 100% bootstrap values. There are a few aspects in which our tree differs from other gene-content trees, however. We have compared our result to that of Wolf et al. (2002). In their study, the genome distance between species (*A* and *B*) was calculated $D_{AB} = 1 - J_{AB}$, where J_{AB} is the Jaccard coefficient, which reflects the similarity of gene content between *A* and *B*. Consider the phylogeny of Archaea, for instance. Both studies support that Hbs (*Halobacterium* sp) appears at the root of the tree, and that the Euryachaeota (Afu, Mja, Mth, and Pho; see fig. 3 for species abbreviations) are clustered together. However, our genome phylogeny suggests that the Crenarchaeota "Ape" (*Aeropyrum pernix*) may also branch-off, whereas Wolf et al. (2002) showed that it was clustered with the Euryachaeota Tac (*Thermoplasma acidophilum*). Though it requires further investigation, the genome distance measure used by Wolf et al. (2002) is unlikely additive, so the theoretical basis of their genome tree remains open to question. Indeed, our simulation study has shown that an ad hoc (non-additive) genome distance could be misleading under the "Felsenstein zone" (not shown).

Discussion

Individual gene families may have different phylogenetic trees because of orthology problems caused by fast evolution—gene/genome duplication, or lateral gene transfer (Doolittle 1999*b*, Eisen 2000; Gu, Wang, and Gu 2002; Jordan et al. 2001; Gu and Huang 2002). The

Table 2
Correct Percentage (%) of Tree Making: Unequal Branch Lengths (see fig. 2B C).

	The c/a_l Ratio				
N	1	0.8	0.4	0.2	0.1
(1) Refer to fig. 2 B					
100	73	66	58	41	30
500	98	92	80	50	40
2,000	100	100	95	87	78
(2) Refer to fig. 2 C					
100	79	78	75	66	60
500	100	97	92	76	78
2,000	100	100	100	96	95

NOTE.—genome branch lengths: $a_l = 0.6$, $a_s = 0.06$, and $P/L = 0.5$.

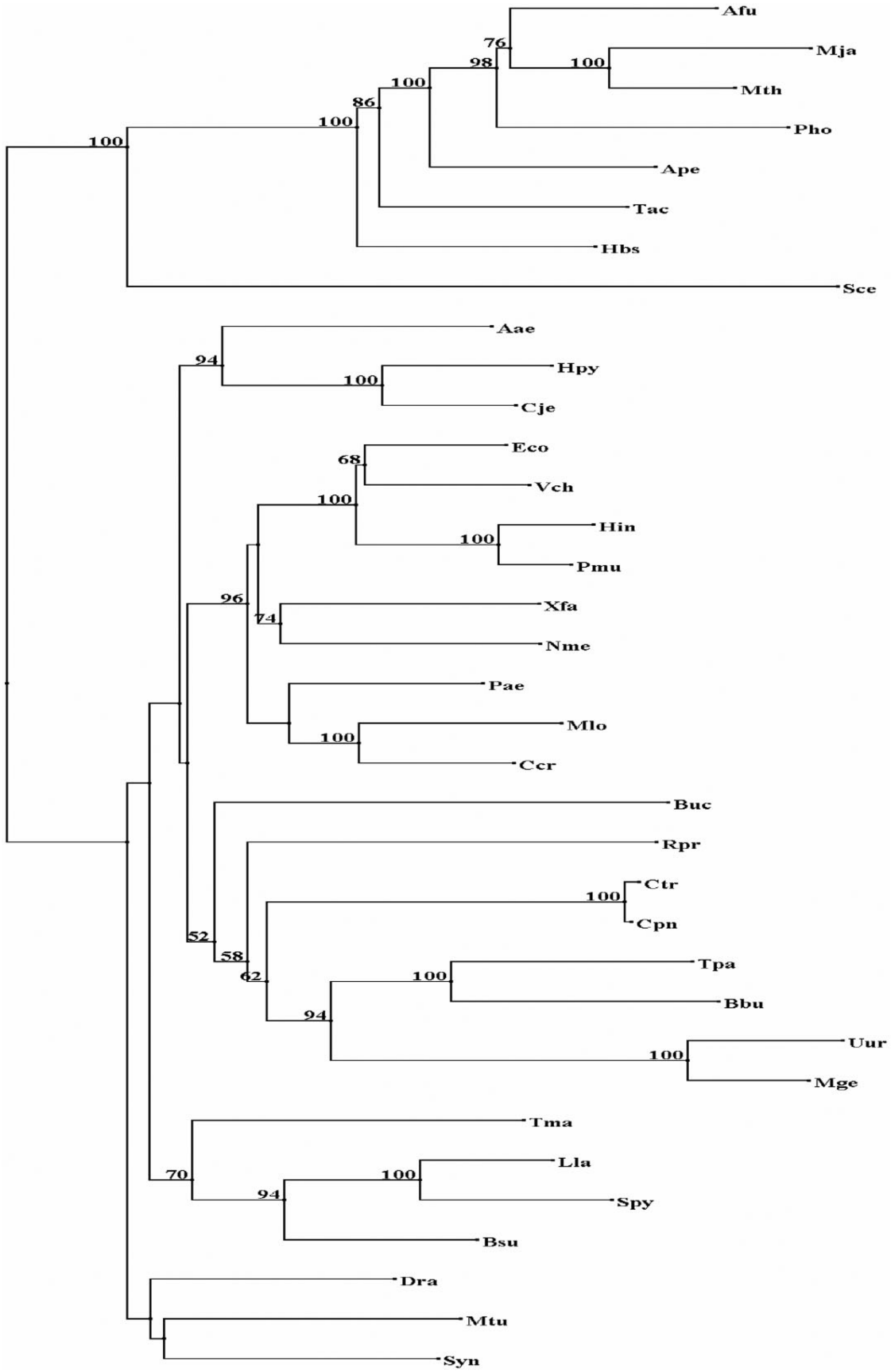

FIG. 3.—The genome phylogeny of 35 microbial complete genomes, inferred by the extended gene-content data set. Bootstrapping values <50% are not presented. Species abbreviations: Archaea: Afu, *Archaeoglobus fulgidus*; Hbs, *Halobacterium sp. NRC-1*; Mja, *Methanococcus jannaschii*; Mth, *Methanothermobacter thermautotrophicus*; Tac, *Thermoplasma acidophilum*; Pho, *Pyrococcus horikoshii*; Ape, *Aeropyrum pernix*. Eukaryota: Sce, *Saccharomyces cerevisiae*. Bacteria: Aae, *Aquifex aeolicus*; Tma, *Thermotoga maritime*; Dra, *Deinococcus radiodurans*; Mtu, *Mycobacterium tuberculosis H37Rv*; Lla, *Lactococcus lactis*; Spy, *Streptococcus pyogenes M1 GAS*; Bsu, *Bacillus subtilis*; Syn, *Synechocystissp.*; Eco, *Escherichia coliK12*; Buc, *Buchnera sp. APS*; Vch, *Vibrio cholerae*; Pae, *Pseudomonas aeruginosa*; Hin, *Haemophilus influenzae*; Pmu, *Pasteurella multocida*; Xfa, *Xylella fastidiosa 9a5c*; Nme, *Neisseria meningitidis MC58*; Hpy, *Helicobacter pylori 26695*; Cje, *Campylobacter jejuni*; Mlo, *Mesorhizobium*

whole-genome approach provides one feasible solution for overcoming this problem. Other methods, including merging individual trees to a biologically meaningful phylogeny, or concatenating well-selected proteins to make a single phylogeny, are certainly also valuable.

We developed a stochastic model for genome evolution under a given phylogeny. However, we have found that it is difficult use the widely cited gene-content data to estimate the additive genome distance. We solved this problem by using the extended gene contents that take duplicate genes into account. Computer simulation shows that the genome phylogeny inference is efficient, consistent, and fairly robust. Moreover, the example of 35 microbial complete genomes demonstrates that the new method is useful not only to study the universal tree of life but also to explore the evolutionary pattern of genomes.

Though many reports of lateral gene transfer (Doolittle and Logsdon 1998; Lawrence and Ochman 1998) have made popular the view that it must be one of the "major forces," at the genome-level, there may be only a small portion of gene families that could be affected. Lateral gene transfer from one organism to another may only increase the size of an existing gene family (type A) in the host genome, or it may introduce new genes into the host genome (type B) (Snel, Bork, and Huynen 1999; Eisen 2000; Sankoff 2001). Our simulation study has shown that the genome tree is virtually unaffected by type A lateral gene transfer, and not very sensitive to type B lateral gene transfer except when it is overwhelming (unpublished result). Although the relative contributions of these two types of lateral gene transfer is yet to be determined, the genome tree seems to be robust against lateral gene transfer. Indeed, our example shows the correspondence of the genome tree (fig. 3) with the 16s rRNA tree (Snel, Bork, and Huynen 1999). Further study will show whether the genome tree can be used as an "independent" phylogenetic framework upon which to construct and test evolutionary hypotheses, including the pattern of lateral gene transfer.

Further studies should take two directions. The first one is to improve the evolutionary model. For instance, the evolutionary rates of gene proliferation or gene loss (λ and μ) could vary not only among gene families but also among lineages (Aravind et al. 2000). One may try some techniques (Gu, Fu, and Li 1995; Gu 1999) developed for sequence evolution to relax the assumption of constant rate. All gene-content–based methods actually assume independent evolution of gene families, which may not be realistic. Because gene families within similar metabolic pathways may tend to co-evolve (Pellegrini et al. 1999); that is, their presence/absence may not be independent among gene families, we shall study this problem under the phylogenetic framework in the future. It remains a challenge to find ways to model the effect of lateral gene transfer. The second direction for future studies involves means of implementing more sophisticated tree-making algorithms. We shall develop some fast but heuristic algorithms so that the ML phylogeny can be used in practice. The Bayesian inference in phylogenetics is also worth considering, though the controversy remains unresolved (Huelsenbeck et al. 2001; Susuki, Glazko, and Nei 2002; Alfaro, Zoller, and Lutzoni 2003).

Acknowledgments

This work was supported by National Institutes of Health grant number RO 1 GM 62118. The computer program is available at the Web site http://xgu.zool.iastate.edu.

Literature Cited

Alfaro, M. E., S. Zoller, and F. Lutzoni. 2003. Bayes or Bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. **20**:255–266.

Aravind, L., H. Watanabe, D. J. Lipman, and E. V. Koonin. 2000. Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc. Natl. Acad. Sci. USA **97**:11319–11324.

Clarke, G. D. P., R. G. Beiko, M. A. Ragan, and R. L. Charlebois. 2002. Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores. J. Bacteriol. **184**:2072–2080.

Daubin, V., N. A. Moran, and H. Ochman. 2003. Phylogenetics and the cohesion of bacterial genomes. Science **301**:829–832.

Doolittle, W. F. 1999*a*. Phylogenetic classification and the universal tree. Science **284**:2124–2129.

Doolittle, W. F. 1999*b*. Technical comments (Response) on Doolittle (1999*a*). Science **286**:1443a.

Doolittle, W. F., and J. M. Logsdon. 1998. Archaeal genomics: do Archaea have a mixed heritage? Curr. Biol. **8**:R209–R211.

Eisen, J. A. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr. Opin. Genet. Dev. **10**:606–611.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

Fitz-Gibbon, S. T., and C. H. House. 1999. Whole genome–based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. **27**:4218–4222.

Golding, G. B., and R. S. Gupta. 1995. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol. Biol. Evol. **12**:1–6.

Gu, X., Y. X. Fu, and W. H. Li. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. **12**:546–557.

Gu, X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. **16**:1664–1674.

Gu, X. 2000. A simple evolutionary model for genome phylogeny inference based on gene content. Pp. 515–524 *in* D. Sankoff and J. H. Nadeau, eds. Comparative genomics. Kluwer Academic Publishers, Dordrect, The Netherlands.

Gu, X. 2001. Maximum likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. **18**:453–464.

Gu, X., and W. Huang. 2002. Testing the parsimony test of genome duplications: a counterexample. Genome Res. **12**:1–2.

←

loti; Ccr, *Caulobacter crescentus*; Rpr, *Rickettsia prowazekii*; Ctr, *Chlamydia trachomatis*; Cpn, *Chlamydophila pneumoniae*; Tpa, *Treponema pallidum*; Bbu, *Borrelia burgdorferi*; Uur, *Ureaplasma urealyticum*; Mge, *Mycoplasma genitalium.*

Gu, X., Y. Wang, and J. Gu. 2002. Age-distribution of human gene families showing equal roles of large and small-scale duplications in vertebrate evolution. Nature Genet. **31**: 205–209.

House, C. H., and S. T. Fitz-Gibbon. 2002. Using homolog groups to create a whole-genomic tree of free-living organisms: an update. J. Mol. Evol. **54**:539–547.

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294**:2310–2314.

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

Huynen, M. A., and B. Snel. 2000. Gene and context: integrative approaches to genome analysis. Adv. Prot. Chem. **54**:345–379.

Jordan, I. K., K. S. Makarova, J. L. Spouge, Y. I. Wolf, and E. V. Koonin. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. Genome Res. **11**:555–565.

Korbel, J. O., B. Snel, M. A. Huynen, and P. Bork. 2002. SHOT: a Web server for the construction of genome phylogenies. Trends Genet. **18**:158–62.

Lawrence, J. G., and H. Ochman. 1998. Molecular archaeology of the Escherichia coli genome. Proc. Natl. Acad. Sci. USA **95**:9413–9417.

Lin, J., and M. Gerstein. 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res. **10**:808–818.

Natale, D. A., U. T. Shankavaram, M. Y. Galperin, Y. I. Wolf, L. Aravind, and E. V. Koonin. 2000. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). Genome Biol. 1, RESEARCH0009.

Nei, M., X. Gu, and R. Sitnikova. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA **94**:7799–7806.

Nelson, K. E., R. A. Clayton, S. R. Gill, et al. (29 co-authors). 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature **339**:323–329.

Olsen, G. J., C. R. Woese, and R. Overbeek. 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. **176**:1–6.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA **96**:4285–4288.

Saitou, N., and M. Nei. 1987. The Neighbor-Joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

Sankoff, D. 2001. Gene and genome duplication. Curr. Opin. Genet. Dev. **11**:681–684.

Snel, B., P. Bork, and M. A. Huynen. 1999. Genome phylogeny based on gene content. Nat. Genet. **21**:108–110.

Suzuki, Y., G. V. Glazko, and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proc. Natl. Acad. Sci. USA **99**:16138–16143.

Tekaia, F., A. Lazcano, and B. Dujon. 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. **9**:550–557.

Woese, C. 1998. The universal ancestor. Proc. Natl. Acad. Sci. USA **95**:6854–6859.

Wolf, Y., I. B. Rogozin, N. V. Grishin, and E. V. Koonin. 2002. Genome trees and the tree of life, Trends Genet. **18**:472–479.

Brian Golding, Associate Editor

Accepted March 18, 2004