Personalized Monograph

2023v1

Genome Phylogeny and the Tree of Life

Xun Gu

Unit-1

Genome phylogenetic analysis based on extended gene contents

Unit-2

GeneContent: software for whole-genome phylogenetic analysis

Unit-1

Genome Phylogenetic Analysis Based on Extended Gene Contents

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, {\em Haemophilus influenzae}. Up 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 for having a meaningful interpretation of the Tree of Life. Because phylogenetic analysis based on single-gene (family) has produced many conflicted gene trees, the long-term controversy between ``vertical'' (tree-like) evolution and lateral (horizontal) gene transfer is being more heated rather than resolved in the genome era (Golding and Gupta 1995; Doolittle and Logsdon 1998; Jain et al. 1998; Doolittle 1999a, 1999b; Huynen and Snel 2000; Nelson et al. 1999; Tekaia et al. 1999; Wolf et al. 2002; Daubin et al. 2003).

Since 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 extract the bulk phylogenetic signals. 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 {\em 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 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 (Doolittle 1999b) raised a fundamental question whether a genome tree based on gene content is only the best phenotypic measure, rather than the evolutionary relationship. 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 et al. 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 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 dominate picture of evolution (Wolf et al. 2002).

We have recognized the important role of modeling for phylogenomic analysis for justifying whether the inferred tree indeed represents the genome phylogeny. Since 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.

Joint Size Distribution of the Gene Family in Multiple Genomes

The whole-genome comparison has revealed a high variation of the size of gene families among complete genomes, because a gene family can be generated, 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 gene family: gene loss (non-functionalization 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 to be 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

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

where the proliferation parameter α and the loss parameter 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. Eq.(2) implies $\alpha/\beta = \lambda/\mu$, as called the *P/L* ratio. For the size of gene family under the birth-death model is expected to be $X_0 e^{-(\lambda-\mu)t}$. It appears that $\alpha > \beta$ (or *P/L*>1) indicates, on average, the increase of gene family size during the evolution and *vice versa*.

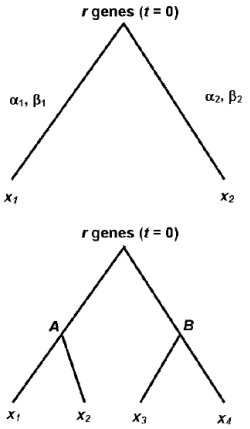

FIG. 1.—Schematic genome evolution for two genomes and four genomes, respectively. The gene family has r member genes in the root. After t evolutionary time units, the size of the gene family is x_1 and x_2 in genomes 1 and 2, respectively. For four genomes, the size of the gene family is x_i (i =1,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

Consider two genomes that have been diverged t time units ago (Fig.1). For a given gene family, assume that there are r member genes at t=0 (in the common ancestor), and X_i number genes in each genome i=1, 2, respectively. Under the assumption of independent evolution between lineages, the (conditional) joint probability is given by

$P(X_1, X_2|X_0=r)=P(X_1|X_0=r)\times P(X_2|X_0=r)$. Since 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

$$P(X_1, X_2) = \sum_{r=1}^{\infty} \pi(r) P(X_1, X_2|X_0 = r) = \sum_{r=1}^{\infty} \pi(r) P(X_1|r) P(X_2|r) \quad (3)$$

For the general n-genomes, let X_i represent the size of a gene family in the i-th genome, i=1,…, n. The joint size distribution of the gene family $X=(X_1, ..., X_n)$ 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

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

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

Two-Genome Model and Expression Distances

Given the joint-size distribution, say, Eq.(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 at first one should define an additive genome distance measure. With some algebras from Eq.(2), two quantities, the proliferation measure d_λ and the loss measure d_μ, are given by

$$\begin{aligned} d_\lambda &= \frac{\alpha}{\beta - \alpha} \ln \frac{1-\alpha}{1-\beta} = \lambda t \\ d_\mu &= \frac{\beta}{\beta - \alpha} \ln \frac{1-\alpha}{1-\beta} = \mu t \end{aligned} \quad (5)$$

respectively. For two genomes (Fig.1), let λ_i, μ_i, α_i, β_i, $d_\lambda(i)$ and $d_\mu(i)$ be the corresponding parameters in each lineage, i=1, 2; see Eqs.(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 Eq.(5) it is given by

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

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

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

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

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

Apparently, these genome distance measures are additive and $G_P/G_P=P/L$ ratio. Eqs.(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., Tekaia et al 1999; Snel et al. 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 i-th genome; otherwise Y_i=0. Therefore, gene content pattern is the most degenerated size distribution of the gene family. In the following we will show that it becomes insufficient for estimating the genome distance.

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

$$P(Y_1, Y_2) = \sum_{r=1}^{\infty} \pi(r) P(Y_1|r) P(Y_2|r) \tag{9}$$

Since $P(Y_i=0|r)=\beta_i^r$, and $P(Y_i=1|r)=1-\beta_i^r$, 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 Eq.(9) we have

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

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

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

Since Eq.(10) only relies 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 Eq.(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) 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 Eq.(1), their probabilities are $P(Z=0|X_0=r)=P(X_t=0|X_0=r)$, $P(Z=1|X_0=r)=P(X_t=1|X_0=r)$ and $P(Z=2|X_0=r)=\Sigma_{k\geq 2}P(X_t=k|X_0=r)$, as given by

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

respectively.

The joint-distribution for two genomes

Consider two genomes that have been 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 i-th genome, i=1, 2. Similar to Eq.(3) and Eq.(9), the joint distribution of Z_1 and Z_2 is given by

$$P(Z_1, Z_2) = \sum_{r=1}^{\infty} \pi(r) P(Z_1, Z_2 | X_0 = r) = \sum_{r=1}^{\infty} \pi(r) P(Z_1|r) P(Z_2|r) \quad (13)$$

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

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

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 Eq.(11), the function $R(\beta)$ is given by

$$R(\beta) = \sum_{r=1}^{\infty} \pi(r) r \beta^{r-1} = \frac{f}{1-(1-f)\beta} + \frac{f(1-f)\beta}{[1-(1-f)\beta]^2} \quad (15)$$

and the function $S(\beta)$ is given by

$$S(\beta) = \sum_{r=1}^{\infty} \pi(r) r^2 \beta^{r-1} = \frac{f}{1-(1-f)\beta} + \frac{3f(1-f)\beta}{[1-(1-f)\beta]^2} + \frac{2f(1-f)^2\beta^2}{[1-(1-f)\beta]^3} \quad (16)$$

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-based approach to estimating the genome distances. Usually the prior parameter f can be estimated from the observed size frequencies of gene families. Since 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

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

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

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

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 and β_2) are estimated, the computation of genome distances by Eqs.(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. Because of the space limitation, we 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 parameters α_i and loss parameters β_i (i=1, 2), and then compute the genome distances according to Eqs.(6)-(8). The mean and squared root of variance for each estimate are used for examining the statistical properties.

We have studied four typical cases: the gene-loss model (λ=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 N=200, 500, and 1000, respectively. We have examined a variety of combinations from these models in two lineages, and 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 increasing of the number of gene families, which 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 using the neighbor-joining (NJ) algorithm. The efficiency of phylogenetic inference is then measured by the percentage of correct topology inference over 1000 replicates. After having examined many combinations, we conclude that our method is efficient, that is, except for some extreme cases, the correct percentage is satisfactory (>70%) when N>500, and consistent, that is, the correct percentage tends to be 100% when $N\rightarrow\infty$. When the internal branch length is short, the genome tree inference can be significantly improved as N becomes large. While it is easy to achieve in the case of four equal external branch lengths, it requires much more number of gene families when the external branches are highly unequal. Nevertheless, even in the very extreme case, the correct percentage of tree-making is close to 100% for sufficient 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. After we examined many cases, we found that the performance of tree-making is very robust against the choice of a specific form of $\pi(r)$.

Technical comments

Power of whole-genome approach

Individual gene families may have different phylogenetic trees because of orthology problem, caused by fast evolution, gene/genome duplication, or lateral gene transfer (Doolittle 1999b, Jordan et al 2001; Gu et al. 2002; Eisen 2000; Gu and Huang 2002). The whole-genome approach provides one feasible solution to overcome 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.

Effect of lateral gene transfer

Though many reports of lateral gene transfer (Doolittle and Logsdon 1998, Lawrence and Ochman 1998) have made a popular view that it must be one of ``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 existed gene family (Type A) in the host genome, or may introduce new genes into the host genome (Type B) (Snel et al 1999; Eisen 2000; Sankoff 2001). Our simulation study has shown that the genome tree is virtually not affected by the type A, and not very sensitive to the 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 know, the genome tree seems to be robust against the lateral gene transfer. Indeed, our example that shows the correspondence of the genome tree with the 16s rRNA tree (Snel et al 1999). Further study will show whether the genome tree can be used as ``independent" phylogenetic framework upon which to construct and test evolutionary hypotheses, including the pattern of lateral gene transfer.

Further studies include 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 2000). One may try some techniques (Gu et al 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. Since gene families within similar metabolic pathways may tend to co-evolve (Pellegrini et al. 1999), that is, the 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 how to model the effect of lateral gene transfer. The second direction is how to implement more sophisticated tree-making algorithms. We shall develop some fast but heuristic algorithms so that the maximum likelihood 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 et al. 2002; Alfaro et al. 2003).

Appendix in-addition: a simple algorithm to estimate genome distances

Estimation of loss parameters (β_1 and β_2)

After some algebras from Eq.(14), one can show

$$\frac{P(0,1)+P(0,2)}{1-P(0,0)} = 1 - \frac{1-Q(\beta_1)}{1-Q(\beta_1\beta_2)} \qquad \text{(A-1)}$$

where $Q(\beta)$ is given by Eq.(11). Let N be the number of gene families and $N_0^{(1)}$ be the number of gene families that have no member in the first genome. It appears that the term on the left hand of Eq.(A-1) can be estimated by $N_0^{(1)}/N$. Hence, the first estimation equation can be written as

$$1 - \frac{N_0^{(1)}}{N} = \frac{1-Q(\hat{\beta}_1)}{1-Q(\hat{\beta}_1\hat{\beta}_2)} \qquad \text{(A-2)}$$

In the same manner, the second estimation equation is then given by

$$1 - \frac{N_0^{(2)}}{N} = \frac{1-Q(\hat{\beta}_2)}{1-Q(\hat{\beta}_1\hat{\beta}_2)} \qquad \text{(A-3)}$$

Further, the following relationship is useful to simplify the numerical analysis, that is,

$$\frac{1}{1-\beta_2} = \frac{A}{1-\beta_1} + B \qquad \text{(A-4)}$$

where A=(1- $N_0^{(1)}/N$)/(1-$N_0^{(2)}/N$), and B=(A-1)(1-f)/f.

Estimation of proliferation parameters (α_1 and α_2)

We first focus on the estimation of α_1. From Eq.(14), one can show

$$\frac{P(1,0)}{1-P(0,0)} = \frac{\beta_2 R(\beta_1\beta_2)}{1-Q(\beta_1\beta_2)}\omega_1 \qquad \text{(A-5)}$$

It appears that the term on the left hand of Eq.(A-5) can be estimated by n_{10}/N, where n_{10} is the number of gene families that have zero member in the first genome and the single member in the second genome. The second equation we will consider is

$$\frac{P(1,1)+P(1,2)}{1-P(0,0)} = \left[\frac{R(\beta_1)-\beta_2 R(\beta_1\beta_2)}{1-Q(\beta_1\beta_2)}\right]\omega_1 \qquad \text{(A-6)}$$

It appears that the term on the left hand of Eq.(A-6) can be estimated by $(n_{11}+n_{12})/N$, where n_{11} is the number of gene families that have single member in both genomes and n12 is that where the second genome has multiple members. Together, we have the following objective function

$$J = \left(\frac{n_{10}}{N} - a_1\omega_1\right)^2 + \left(\frac{n_{11}+n_{12}}{N} - b_1\omega_1\right)^2 \qquad \text{(A-7)}$$

where two parameters a_1 and b_1 can be empirically given by

$$\begin{aligned} a_1 &= \frac{\hat{\beta}_2 R(\hat{\beta}_1\hat{\beta}_2)}{1-Q(\hat{\beta}_1\hat{\beta}_2)} \\ b_1 &= \frac{R(\hat{\beta}_1)-\hat{\beta}_2 R(\hat{\beta}_1\hat{\beta}_2)}{1-Q(\hat{\beta}_1\hat{\beta}_2)} \end{aligned} \qquad \text{(A-8)}$$

respectively. We then obtain the estimate of ω_1 by minimizing J, that is,

$$\hat{\omega}_1 = \frac{a_1 n_{10} + b_1(n_{11}+n_{12})}{(a_1^2+b_1^2)N} \qquad \text{(A-9)}$$

In the same manner, we have

$$\hat{\omega}_2 = \frac{a_2 n_{01} + b_2(n_{11}+n_{12})}{(a_2^2+b_2^2)N} \qquad \text{(A-10)}$$

where two parameters a_2 and b_2 are given by

$$\begin{aligned} a_2 &= \frac{\hat{\beta}_1 R(\hat{\beta}_1\hat{\beta}_2)}{1-Q(\hat{\beta}_1\hat{\beta}_2)} \\ b_2 &= \frac{R(\hat{\beta}_2)-\hat{\beta}_1 R(\hat{\beta}_1\hat{\beta}_2)}{1-Q(\hat{\beta}_1\hat{\beta}_2)} \end{aligned} \qquad \text{(A-11)}$$

Noting that ω_k=(1-β_k)(1-α_k), k=1,2, we finally obtain

$$\begin{aligned}\hat{\alpha}_1 &= 1-\frac{\hat{\omega}_1}{1-\hat{\beta}_1}\\ \hat{\alpha}_2 &= 1-\frac{\hat{\omega}_2}{1-\hat{\beta}_2}\end{aligned} \qquad \text{(A-12)}$$

Further reading

Gu,X. and Zhang,H.M. (2004) Genome phylogenetic analysis based on extended gene contents. *Mol. Biol. Evol.*, **21**, 1401–1408.

References

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.

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

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.

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.

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.

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.

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.

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

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.

Unit-2

GeneContent: software for whole-genome phylogenetic analysis

Since phylogenetic trees inferred from individual genes may be inconsistent, the whole-genome approach, such as the gene con- tent, becomes an attractive approach to extract bulk phylogenetic signals. For instance, some authors (e.g. 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).

However, the statistical model of genome evolution should be addressed appropriately for having a reliable phylogenetic inference rather than the best phenotypical clustering. To this end, Gu and Zhang (2004) proposed a statistical framework for the phylogenetic gene-content analysis, which has been successfully applied for the tree of life. We have subsequently developed a user-friendly GUI-based software system, GeneContent, to facilitate the further study in comparative genomics.

The software GeneContent deals with two types of gene-content data: the conventional gene content (Snel *et al*., 1999; Huynen *et al*., 1999; Lin and Gerstein, 2000; Korbel *et al*., 2002) contains the genome-wide information for the presence/absence of gene families across multiple species, while the extended gene content (Gu and Zhang, 2004) contains the genome-wide information as follows: absence of a gene family, presence as single copy or presence as duplicates. The advantage of extended gene content for phylogenomics is demonstrated below.

Based on the birth–death stochastic model (Gu and Zhang, 2004), an additive genome distance measure between two species can be defined as $G=2(\lambda+\mu)t$, where λ is the proliferation (duplicate) rate of a gene family, μ is the loss rate of genes and t is the evolutionary time units. It has been shown that for two genomes, it is difficult to utilize the conventional gene-content data to estimate the genome distance G, except for the special case, where $\lambda = 0$. Gu and Zhang (2004) have solved this problem by introducing the concept of extended gene content, and proposed an efficient algorithm for genome-wide phylogenetic analysis since it does not require much computational time.

The interface of the software GeneContent (Fig. 1) is straightforward and easy to use. The input of the data is in the text file, in which the rows correspond to different genomes and the columns to gene families. The values for each entry of the data matrix could represent the size of gene family in the genome, gene content or extended gene content. Our program will trim the input matrix to fit the type of input as specified by the user. GeneContent provides three options to calculate genome distance: the Poisson distance, the gene content (under the special case where $\lambda=0$) and the extended gene content. By default, both gene content and extended gene content methods will be provided, except that the input matrix only contains two types of values (i.e. 0 for absence and 1 for presence); in this case, the extended gene content method will be disabled. The Poisson distance is available for comparison purpose. Note that the gene-content dis- tance between species (A and B) is calculated $D_{AB}=1-J_{AB}$, where J_{AB} is the Jaccard coefficient, which reflects the similarity of gene content between A and B (Wolf *et al*., 2002).

After obtaining the genome distance matrix, the software is able to infer the genome phylogeny using the neighbor-joining method (Saitou and Nei, 1987). The statistical reliability of the inferred genome phylogeny is examined by the conventional bootstrapping approach. Since the inferred phylogeny is un-rooted, the option for changing the root under the tree-view is available, as well as other options for visualization editing. The inferred genome tree can be saved as a text file in the Phylip format, which is useful in some cases.

The performance of the above algorithm has been examined by the universal genome tree of 36 complete genomes (Gu and Zhang, 2004). In the current version, we have implemented some options to explore the pattern of genome evolution. For instance, the proliferation/loss rate ratio can be mapped onto the phylogenetic tree and the bootstrapping test can be performed to examine whether it remains a constant among lineages. We will upgrade our software in two directions. The first one is to improve the evolutionary model by considering more

factors such as lateral gene transfer and co-evolution among gene families. The second direction is to implement more sophisticated tree-making algorithms, e.g. a fast algorithm for the maximum-likelihood inference of genome phylogeny.

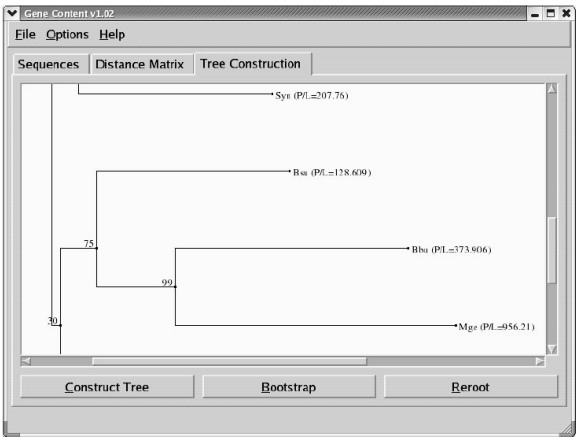

Fig. 1. The main interface of GeneContent includes three tabs: sequences, distance matrix and tree construction

Further reading

Gu, X*, Huang, W, Xu, D, Zhang H (2005) GeneContent: Software for Whole-Genome Phylogenetic Analysis. Bioinformatics 21(8):1713-1714.

REFERENCES

Clarke,G.D.P., Beiko,R.G., Ragan,M.A. and Charlebois,R.L. (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.

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

Gu,X. and Zhang,H.M. (2004) Genome phylogenetic analysis based on extended gene contents. *Mol. Biol. Evol.*, **21**, 1401–1408.

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

Huynen,M.A., Snel,B. and Bork,P. (1999) Technical comments on Doolittle [1999a].

Science, **286**, 1443a.

Korbel,J.O., Snel,B., Huynene,M.A. and Bork,P. (2002) SHOT: a web server for the construction of genome phylogenies. *Trends Genet.* **18**, 158–162.

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

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

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for recon- structing phylogenetic trees. *Mol. Bol. Evol.*, **4**, 406–425.

Snel,B., Bork,P. and Huynen,M.A. (1999) Genome phylogeny based on gene content.

Nat. Genet., **21**, 108–110.

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