Statistical Methods for Testing Functional Divergence after Gene Duplication

Xun Gu

Department of Zoology/Genetics, Iowa Computational Molecular Biology Laboratory, Iowa State University

Functional innovations after gene duplication may result in altered functional constraints between member gene clusters of a gene family. This type (type I) of functional divergence is measured by the coefficient of functional divergence (θ_λ), which can be interpreted as the decrease in rate correlation between gene clusters, or the probability that the evolutionary rate at a site is statistically independent between two gene clusters. A simple stochastic model has been developed for estimating θ_λ and testing its statistical significance. The current model includes the model of rate variation among sites as a special case when $\theta_\lambda = 0$. Moreover, we have developed a site-specific profile based on the hidden Markov model to identify critical amino acid residues that are responsible for these functional differences between two gene clusters, which may have great potential in functional genomics.

Introduction

An understanding of the functional diversity of a gene family has been a major component in molecular evolutionary study (Nei 1987; Li 1997). Recently, its importance for functional genomics has been well recognized (Henikoff et al. 1997; Bork and Koonin 1998). Indeed, many organisms have undergone genomewide or local chromosome duplication events during their evolution (Ohno 1970; Lundin 1993; Holland et al. 1994; Spring 1997). Moreover, new types of multiple-domain proteins can be generated by the domain-shuffling mechanism (Henikoff et al. 1997). As a consequence of these gene/genome duplication and domain-shuffling events, many genes are represented as several paralogs in the genome with related but distinct functions. These gene family proliferations are thought to have provided the raw materials for functional innovations (Li 1983; Nei 1987; Lundin 1993; Hughes 1994; Henikoff et al. 1997). It has been widely accepted that following gene duplication, one gene copy maintains the original function, while the other copy is free to accumulate amino acid changes as a result of functional redundancy or positive selection (Li 1983). Unless this type of functional divergence results in some new functions, over time all but one gene copy will be silenced by deleterious mutations.

Extensive studies have been reported on the underlying mechanism of functional divergence after gene duplication (e.g., Kimura and Ota 1974; Li 1983; Nei 1987; Zhang, Rosenberg, and Nei 1998). Hughes (1994) speculated that the ancestral gene might already be bifunctional and gene duplication simply allows each copy to specialize for one of several functions. Having realized the importance of coevolution between the interacted molecules (e.g., ligand/receptor), Fryxell (1996) argued that functional divergence may occur only when all genes in a pathway are duplicated simultaneously, e.g., by a genome duplication. Nevertheless, it becomes clear that some evolutionary changes in the coding and/or regulatory regions after gene duplication must be responsible for the functional differences between members of a gene family.

An interesting question is whether we can identify these important amino acid (or nucleotide) sites; the methods for doing so may have great potential for functional genomics since they are extremely cost-effective, and the predictions obtained can be further tested by experimentation (Golding and Dean 1998). For example, we may infer amino acid sites that have experienced altered functional roles in a period of evolution. Since amino acid differences between two gene family members can be the result of either an ancient gene duplication or a more recent event that is subject to rapid functional divergence, a homologous search based on sequence similarity (score) may not be sufficient for our purpose (Eisen 1998; Golding and Dean 1998). In fact, a simple application of molecular phylogenetic analyses cannot completely distinguish between these possibilities. Moreover, many observed amino acid changes are due to purifying selection and are not directly related to functional innovations. This problem is serious in practice because experimental evidence from many case studies has already shown that functional divergence after gene duplication can be generated by only a few amino acid changes (for a review, see Golding and Dean 1998).

The purpose of this paper is to develop a novel stochastic model for functional divergence after gene duplication, to estimate the level of functional divergence, and to predict important amino acid residues for these functional differences between member genes of a gene family. We shall develop a quantitative measure for the functional difference that can be estimated from sequence data and distinguish between these changes related to functional divergence and the background changes which mainly represent neutral evolution. The present method is applied to the transferrin and myc gene families to demonstrate its potential in functional and comparative genomics.

Functional Divergence and Altered Functional Constraint

A (homologous) gene cluster is defined as a monophyletic group of sequences under a phylogenetic tree.

Key words: gene family evolution, gene duplication, functional divergence, hidden Markov model, site-specific profile.

Address for correspondence and reprints: Xun Gu, Department of Zoology/Genetics, 332 Science II Hall, Iowa State University, Ames, Iowa 50011. E-mail: xgu@iastate.edu.

Mol. Biol. Evol. 16(12):1664–1674. 1999

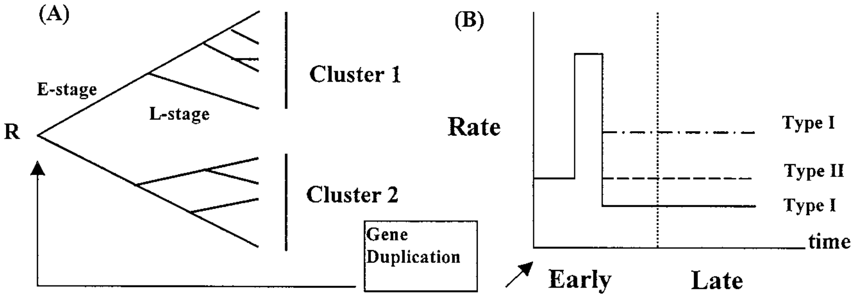

FIG. 1.—*A*, Two gene clusters after gene duplication; E and L represent early and late stages of gene cluster 1, respectively. *B*, Type I and type II functional divergences after gene duplication. In the early stage, the evolutionary rate (say, in cluster 1) may increase for functional-divergence-related change, but in the late stage, it may be higher (or lower) than the original rate, resulting in altered functional constraints between clusters 1 and 2 (type I functional divergence). If the rate in the late stage is the same as the original one again, no altered functional constraints between clusters 1 and 2 can be observed (type II functional divergence).

For example, two gene clusters are generated by an event of gene duplication, and each of them consists of several orthologous sequences (fig. 1*A*). It is commonly believed (Li 1983) that after gene duplication, the evolutionary rate (λ) at an amino acid site may increase and functional divergence may occur in the early stage, followed by the late stage, in which purifying selection plays a major role in maintaining related but distinct functions (fig. 1*B*). The underlying mechanism for this type of accelerated evolution after gene duplication is still in dispute (e.g., Li 1983; Nei, Gu, and Sitnikova 1997). If the early-stage functional divergence occurred in one duplicate gene, changes of functional roles at the sites involved can be observed in the late stage. As a result, evolutionary rates at these sites are different between the two gene clusters. Such functional divergence, resulting in altered functional constraint, is called type I functional divergence.

The central tenet of our approach is that type I functional divergence after gene duplication is highly correlated with the change in evolutionary rate, which is analogous to a fundamental rule in molecular evolution: functional importance is highly correlated with evolutionary conservation (Kimura 1983). Alternatively, type II functional divergence does not result in different functional constraints between the two gene clusters, but evolutionary rates can be different between early and late stages (fig. 1*B*). For example, cluster-specific residues may be subject to this type of functional divergence. In this paper, we deal mainly with type I functional divergence; type II functional divergence will be discussed elsewhere. The relationship between functional divergence, altered functional constraint, and evolutionary rate provides a theoretical basis for modeling the type I functional divergence during sequence evolution.

A Simple "Model-Free" Method

Rate Correlation Between Two Gene Clusters

If all sites have experienced no functional divergence after gene duplication, the two duplicate genes have no altered functional constraints, so the evolutionary rate of a site is always the same (or proportional) between them, i.e., the coefficient of rate correlation (over sites) is 1. Obviously, altered functional constraints caused by functional divergence will reduce the rate correlation. Consider a multiple alignment of amino acid sequences containing two gene family members (fig. 1). If orthologous sequences are functionally equivalent, the evolutionary rate (λ) of a site remains constant (or proportional) among branches within a gene cluster, although it may vary among sites. Since a molecular clock is not assumed, lineage-specific factors such as generation time effect (Wu and Li 1985; Gu and Li 1992) will not affect our results. Hence, without loss of generality, the evolutionary rates in gene cluster 1 and gene cluster 2 are simply denoted by λ_1 and λ_2, respectively. The altered functional constraints between two gene clusters can be measured by the coefficient of rate correlation between λ_1 and λ_2,

$$r_\lambda = \frac{\mathrm{Cov}(\lambda_1, \lambda_2)}{\sqrt{\mathrm{Var}(\lambda_1)\mathrm{Var}(\lambda_2)}}, \tag{1}$$

where $\mathrm{Var}(\lambda_1)$, $\mathrm{Var}(\lambda_2)$, and $\mathrm{Cov}(\lambda_1, \lambda_2)$ are the variances and covariance of λ_1 and λ_2, respectively. If there is no functional divergence after gene duplication, $r_\lambda = 1$; otherwise, $r_\lambda < 1$. Therefore, a convenient measure for functional divergence can be simply defined as

$$\theta_\lambda = 1 - r_\lambda. \tag{2}$$

As θ_λ increases from 0 to 1, the functional divergence increases from very weak to extremely strong. In this sense, θ_λ is called the coefficient of functional divergence.

The Poisson Model for Amino Acid Substitutions

To avoid confusion, we mention that the term "model-free" means that there is no specific model for rate variation among sites and rate correlation between gene clusters; the method does require a model for amino acid changes at a site. A simple model is the Poisson process: at a given site, the number of amino acid changes (X_i, $i = 1, 2$ for gene clusters 1 and 2, respectively) follows a Poisson distribution, i.e., the probability of $X_i = k$ is given by

$$p_i(k) = \frac{(\lambda_i T_i)^k}{k!} e^{-\lambda_i T_i}, \qquad i = 1, 2, \tag{3}$$

where T_1 and T_2 are the total evolutionary times of clusters 1 and 2, respectively. In section A.1 of the appendix, we show that the coefficient of functional divergence $\theta_\lambda = 1 - r_\lambda$ is given by

$$\theta_\lambda = 1 - \frac{\sigma_{12}}{\sqrt{(V_1 - D_1)(V_2 - D_2)}}, \tag{4}$$

where D_1 and V_1 (or D_2 and V_2) are the mean and variance of the number of changes (over sites) in cluster 1 (or cluster 2), respectively, and σ_{12} is the covariance (over sites) between them.

To estimate θ_λ for equation (4), we need to know the number of changes at each site for each gene cluster (i.e., X_1 and X_2). Since X_1 and X_2 cannot be directly observed from the sequence data, a conventional solution is to use the minimum number of required changes (m) as an approximation, which can be inferred by the parsimony under a known phylogenetic tree (Fitch 1971). However, m is a biased ‘‘estimate’’ for the true number of changes because it does not consider the possibility of multiple hits (Wakeley 1993). This problem has been solved by using a combination of ancestral-sequence inference and maximum-likelihood estimation (Gu and Zhang 1997). Given a phylogeny, Gu and Zhang (1997) have shown that the expected number of changes (X) at a given site is the nonnegative solution of the likelihood equation

$$\sum_{i=1}^{M} \frac{\delta_i b_i}{1 - e^{-\hat{X} b_i/B}} = 1, \tag{5}$$

where B is the total branch length of the gene cluster, and b_i is the ith branch length, $i = 1, \ldots, M$ (M is the total number of branches); $\delta_i = 1$ if there is an amino acid change in the ith branch, otherwise $\delta_i = 0$. Extensive computer simulation has shown that the estimate of mean of expected number of changes, as well as that of variance, is asymptotically unbiased and robust against the accuracy of ancestral amino acid inference. Two interesting special cases are (1) $\hat{X} \approx m$ for short branch lengths, and (2) $\hat{X} = -M \ln(1 - m/M)$ for equal branch lengths.

Statistical Testing

When the numbers of changes at each site in both clusters (X_1 and X_2) are obtained by Gu and Zhang's (1997) method, estimation of θ_λ is simple according to equation (4). Since $\theta_\lambda > 0$ provides evidence for functional divergence after gene duplication, we have to test for statistical significance. Let r_X be the coefficient of correlation between X_1 and X_2, which is defined by

$$r_X = \frac{\sigma_{12}}{\sqrt{V_1 V_2}}. \tag{6}$$

Since r_X reaches its maximum value r_M when $\theta_\lambda = 0$, i.e.,

$$r_X \le r_M = \sqrt{(1 - D_1/V_1)(1 - D_2/V_2)} \tag{7}$$

(see eq. A.9 in the appendix), the null hypothesis H_0: $\theta_\lambda = 0$ is equivalent to $r_X = r_M$. As a standard coefficient of correlation, Fisher's transformation can be used to compute the confidence level of r_X:

$$z = 0.5 \ln\left(\frac{1 + r}{1 - r}\right).$$

Let z_X and z_M, respectively, be the transforms of r_X and r_M. The sampling variance of z_X is approximately $V(z_X) = 1/(N - 3)$, where N is the sequence length. Under the null hypothesis ($r_X = r_M$), the Z score ($Z = (z_X - z_M)\sqrt{N - 3}$) approximately follows a normal distribution. For example, if the Z score is $|Z| > 1.96$, the null hypothesis $\theta_\lambda = 0$ can be rejected at the 5% significance level. Besides, by the delta method, the approximate sampling variance of $\hat{\theta}_\lambda$ can be computed as

$$\mathrm{Var}(\hat{\theta}_\lambda) \approx \frac{1}{N - 3}\left(\frac{1 - r_X^2}{r_M}\right)^2. \tag{8}$$

We should note that although r_X is negatively correlated with θ_λ and useful for constructing a statistical test, it is not a good measure of the level of functional divergence because it is evolutionarily time-dependent (see eq. A.14 in the appendix).

Examples

Transferrins are iron-binding transport proteins which can bind two atoms of ferric iron Fe^{3+}. They are responsible for the transport of iron from sites of absorption and heme degradation to those of storage and utilization. There is only one transferrin-encoding gene in nonmammalian vertebrates such as birds, frogs, and fishes. In mammals, two closely linked tissue-specific genes (3q21–23 in humans) are found which encode serum transferrin (TF) and lactotransferrin (LTF), respectively. Figure 2 shows the phylogenetic tree of the transferrin gene family by the neighbor-joining method (Saitou and Nei 1987); parsimony and likelihood methods give essentially the same topology (results not shown). TF and LTF form separate gene clusters with a high bootstrap (100%) value. Apparently, this gene duplication occurred before the radiation of mammals but after the divergence between birds and mammals.

Based on the phylogenetic tree (fig. 2), all TF/LTF sequences can be divided into three gene clusters: TF, LTF, and genes from nonmammalian vertebrates (vTF). The expected number of changes at each site in each gene cluster was obtained by Gu and Zhang's (1997) method (fig. 3). Figure 4 shows the correlation between the numbers of changes between TF (X_1) and LTF (X_2); the coefficient of correlation is $r_X = 0.37$. The coefficient of functional divergence between each pair of gene clusters is presented in table 1, e.g., $\theta_\lambda = 0.26$ between TF and LTF, which is significantly larger than 0 ($P < 10^{-4}$). Interestingly, the coefficient of functional divergence between TF and LTF is higher than that between TF and vTF ($\theta_\lambda = 0.13$) or between LTF and vTF ($\theta_\lambda = 0.07$), although the average sequence similarity between vTF and TF (as well as between vTF and LTF) is lower than that between TF and LTF. This result implies that functional divergence had occurred after gene duplication in the lineage leading to mammals, resulting in altered functional constraints between the two tissue-specific isoforms TF and LTF.

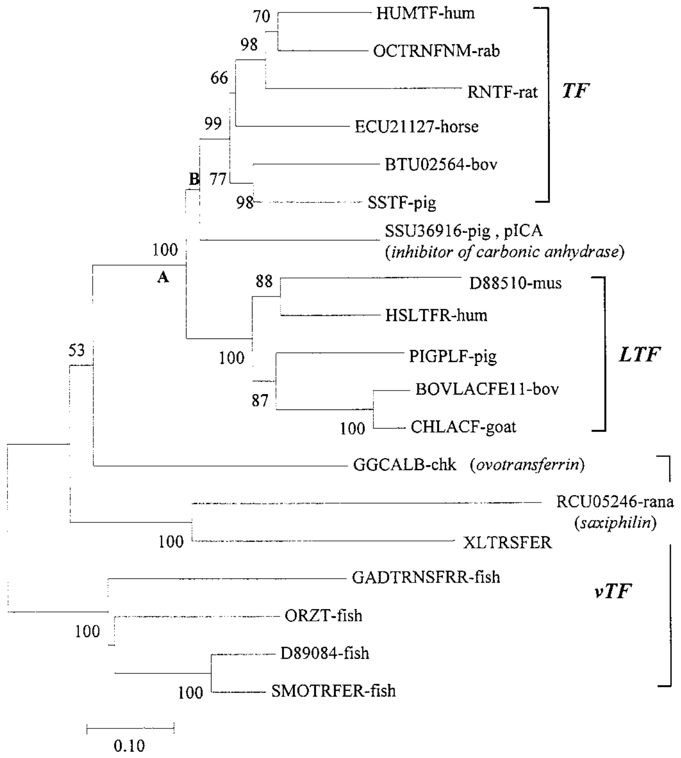

FIG. 2.—The phylogenetic tree of the transferin gene family, which was inferred by the neighbor-joining method using amino acid sequences with Poisson distance. Bootstrapping values >50% are presented.

Two-State Model for Functional Divergence

Although it is simple, the above method requires that each gene cluster should have multiple (say, four) sequences; otherwise, the estimate of θ_λ may be subject to large sampling variance. Therefore, maximum-likelihood (ML) approach is plausible in practice because it has some nice statistical properties.

The Probabilistic Model

Consider an ideal case in which we already know exactly which sites are related to functional divergence. Hence, all sites can be classified into either of two categories, F_0 (functional-divergence-unrelated) or F_1 (functional-divergence-related). In the F_0 category, the evolutionary rate (λ) of a site is the same between gene clusters, indicating no change in functional constraints. In contrast, the evolutionary rate of an F_1 site may have no correlation between gene clusters, because such sites have experienced altered functional constraints. However, in practice, we do not know the category to which each site belongs. This problem is solved by implementing a (two-state) probabilistic model: a given site can be in state F_1 with a probability of $P(F_1)$ or in state F_0 with a probability of $P(F_0)$. Using the same notations as in equation (1), we have $\mathrm{Cov}(\lambda_1, \lambda_2) = P(F_0)\sqrt{\mathrm{Var}(\lambda_1)\mathrm{Var}(\lambda_2)}$, because $\mathrm{Cov}(\lambda_1, \lambda_2 | F_0) = \sqrt{\mathrm{Var}(\lambda_1)\mathrm{Var}(\lambda_2)}$ (completely correlated), and $\mathrm{Cov}(\lambda_1, \lambda_2 | F_1) = 0$ (independent). Then, one can show that

$$P(F_1) = 1 - r_\lambda = \theta_\lambda, \tag{9}$$

where r_λ is the rate correlation between two gene clusters as defined by equation (1). That is, the coefficient of functional divergence (θ_λ) can be interpreted as the probability of a site being in the state of functional divergence (F_1). Denoting the probability of functional divergence at site k by δ_k, we mention that the current two-state model assumes that $\delta_k = 1$ if it is F_1; otherwise, $\delta_k = 0$. Therefore, the proportion of sites expected to be functional-divergence-related is given by $P(F_1) \times 1 + P(F_0) \times 0 = \theta_\lambda$. Furthermore, we assume that the evolutionary rate varies among sites according to a gamma distribution, i.e.,

$$\phi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \tag{10}$$

where $\lambda = \lambda_1$ or λ_2, respectively (Uzzel and Corbin 1971). The shape parameter α describes the degree of rate variation among sites, whereas β is only a scalar. Since $1/\sqrt{\alpha}$ is the coefficient of variation of λ, the larger the α value is, the weaker the rate variation is, and $\alpha = \infty$ means a uniform rate among sites.

The joint distribution of the number of changes, $P(X_1, X_2)$, can be derived as follows. For any F_1 site, the evolutionary rate is statistically independent between two clusters, whereas it is completely correlated at an F_0 site. Thus, the probability of $X_1 = i$ in cluster 1 and $X_2 = j$ in cluster 2 under state F_0 or F_1 is given by

$$P(X_1 = i, X_2 = j | F_1) = Q_1(i)Q_2(j),$$

$$P(X_1 = i, X_2 = j | F_0) = K_{12}(i, j), \tag{11}$$

respectively, where $Q_1(i) = P(X_1 = i | F_1) =$

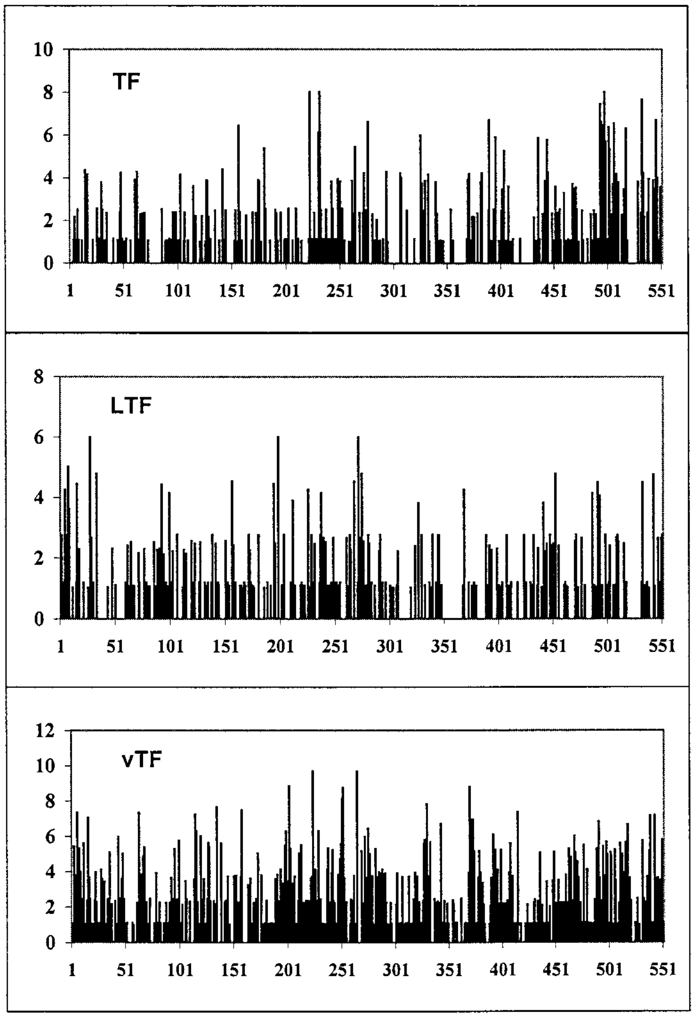

FIG. 3.—The expected numbers of changes at each amino acid site in the TF, LTF, and vTF groups, obtained by Gu and Zhang's (1997) method based on the phylogeny given in figure 2.

$\int_0^\infty p_1(i)\phi(\lambda_1)\, d\lambda_1$, $Q_2(j) = P(X_2 = j \mid F_1) = \int_0^\infty p_2(j)\phi(\lambda_2)\, d\lambda_2$, and $K_{12} = \int_0^\infty p_1(i)p_2(j)\phi(\lambda)\, d\lambda$. It is known that $Q_1(i)$ and $Q_2(j)$ are negative binomial distributions, i.e.,

$$Q_1(i) = \frac{\Gamma(i + \alpha)}{i!\Gamma(\alpha)}\left(\frac{D_1}{D_1 + \alpha}\right)^i\left(\frac{\alpha}{D_1 + \alpha}\right)^\alpha,$$

$$Q_2(j) = \frac{\Gamma(j + \alpha)}{j!\Gamma(\alpha)}\left(\frac{D_2}{D_2 + \alpha}\right)^j\left(\frac{\alpha}{D_2 + \alpha}\right)^\alpha \qquad (12)$$

(Gu and Zhang 1997). After some mathematical simplifications, one can show that $K_{12}(i, j)$ is given by

$$K_{12}(i, j) = \frac{\Gamma(i + j + \alpha)}{i!j!\Gamma(\alpha)}\left(\frac{D_1}{D_1 + D_2 + \alpha}\right)^i\left(\frac{D_2}{D_1 + D_2 + \alpha}\right)^j \times \left(\frac{\alpha}{D_1 + D_2 + \alpha}\right)^\alpha. \qquad (13)$$

Then, the joint distribution is given by $P(X_1, X_2) = P(F_0)P(X_1, X_2 \mid F_0) + P(F_1)P(X_1, X_2 \mid F_1)$, which can be expressed as

$$P(X_1, X_2) = (1 - \theta_\lambda)K_{12} + \theta_\lambda Q_1 Q_2. \qquad (14)$$

One can verify that the joint distribution $P(X_1, X_2)$ has the following properties: (1) The marginal distribution is a negative binomial distribution, i.e.,

$$P(X_1 = i) = \sum_j P(X_1 = i, X_2 = j) = Q_1(i),$$

$$P(X_2 = j) = \sum_i P(X_1 = i, X_2 = j) = Q_2(j), \qquad (15)$$

and (2) the covariance between X_1 and X_2 is given by

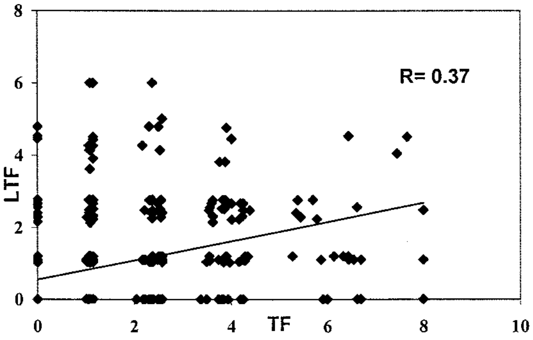

FIG. 4.—The correlation of the numbers of changes between the TF and LTF clusters.

Table 1
Analysis of Functional Divergence Between the TF and LTF Gene Families Based on Equation (4), the Model-Free Estimate

Gene Clusters (1/2)	TF/LTF	TF/vTF	LTF/vTF
D_1	1.17	1.17	0.86
D_2	0.86	2.20	2.20
V_1	2.87	2.87	1.49
V_2	1.49	4.24	4.24
σ_{12}	0.76	1.59	1.04
r_X	0.37	0.46	0.42
r_M	0.50	0.53	0.45
θ_λ	0.26 ± 0.08	0.13 ± 0.06	0.07 ± 0.08
P	$<10^{-3}$	<0.05	>0.10

NOTE.—In the first case, TF represents cluster 1 and LTF represents cluster 2; in the second case, TF represents cluster 1 and vTF represents cluster 2; and in the third case, LTF represents cluster 1 and vTF represents cluster 2. See figure 2 for the definitions of these three clusters. D_1 and V_1 (D_2 and V_2) are the mean and variance of the number of changes in cluster 1 (cluster 2), respectively. σ_{12} is the covariance and r_X is the coefficient of correlation for the numbers of changes between gene clusters 1 and 2. r_M is the expected value of r_X when the evolutionary rate is completely correlated (i.e., $r_\lambda = 1$). The coefficient of rate correlation θ_λ is estimated according to equation (4), and the standard error is given by equation (8). The significance level (P value) is computed by the method of Fisher's transformation.

Table 2
The Coefficients of Functional Divergence (θ_λ) Between Gene Clusters by the Model-Free Method (eq. 4) and the Maximum-Likelihood Method Under the Two-State Model (MLE)

Genes	N	Equation (4)	MLE
TF/LTF	553	0.26 ± 0.08	0.19 ± 0.07
TF/vTF	553	0.13 ± 0.06	0.07 ± 0.03
LTF/vTF	553	0.07 ± 0.08	0.00 ± 0.03
C-myc/N-myc	276	0.52 ± 0.10	0.39 ± 0.08
C-myc/L-myc	276	0.57 ± 0.13	0.56 ± 0.12
N-myc/L-myc	276	0.39 ± 0.12	0.40 ± 0.12

NOTE.—N is the total number of amino acid sites. See figures 2 and 6 for details on each gene cluster.

$$\sigma_{12} = (1 - \theta_\lambda)\frac{D_1 D_2}{\alpha}. \tag{16}$$

When One Gene Cluster Has a Single Sequence

If one cluster (say, cluster 2) has only one single sequence, the joint distribution of X_1 and X_2 needs to be modified, since X_2 has only two states, $X_2 = 0$ or 1, with probabilities $\Pr(X_2 = 0) = e^{-\lambda_2 T_2}$ and $\Pr(X_2 = 1) = 1 - e^{-\lambda_2 T_2}$, respectively. In this case, the joint distribution of X_1 and X_2 at an F_0 site is $P(X_1 = i, X_2 = 0 | F_0) = K_{12}(i, 0)$ and $P(X_1 = i, X_2 = 1 | F_1) = Q_1(i) - K_{12}(i, 0)$. Similarly, the joint distribution of X_1 and X_2 at an F_1 site is $P(X_1 = i, X_2 = 0 | F_0) = Q_1(i)Q_2(0)$ and $P(X_1 = i, X_2 = 1 | F_1) = Q_1(i)[1 - Q_2(0)]$. Then, one can show the joint distribution of X_1 and X_2 as follows:

$$\begin{aligned} P(X_1 = i, X_2 = 0) &= (1 - \theta_\lambda)K_{12}(i, 0) + \theta_\lambda Q_1(i)Q_1(0), \\ P(X_1 = i, X_2 = 1) &= (1 - \theta_\lambda)[Q_1(i) - K_{12}(i, 0)] \\ &\quad + \theta_\lambda Q_1(i)[1 - Q_2(0)]. \end{aligned} \tag{17}$$

Maximum-Likelihood Estimation

Let $P_k(i, j)$ be the probability of $X_1 = i$ and $X_2 = j$ at site k. Thus, the likelihood function can be expressed as

$$L(\mathbf{x} | \text{data}) = \prod_k P_k(X_1 = i, X_2 = j). \tag{18}$$

The parameter set **x** has four parameters, D_1, D_2, α, and θ_λ, which can be numerically estimated by a standard ML approach. Since each marginal distribution follows a negative binomial distribution, we can first use Gu and Zhang's (1997) method for estimating the mean and gamma shape parameter for each gene cluster, i.e., $\hat{D}_1$, $\hat{\alpha}_1$, and $\hat{D}_2$, $\hat{\alpha}_2$. Then, the initial value for α can be simply computed by $\alpha^0 = \sqrt{\alpha_1 \alpha_2}$, and the initial value for θ_λ can be computed by the model-free estimate (eq. 4). Using these initial values, the ML estimates of α and θ_λ, as well as approximate sampling variances, can be obtained numerically (Press et al. 1992). A likelihood ratio test (LRT) is constructed for testing the null hypothesis H_0: $\theta_\lambda = 0$ versus H_A: $\theta_\lambda > 0$. For the likelihood ratio $\text{LR} = \max\{L(H_0 | \text{data})\}/\max\{L(H_A | \text{data})\}$, it is known that $\delta = -2\ \ln(\text{LR})$ asymptotically follows a $\chi^2_{[1]}$. Some examples for ML estimation (MLE) are shown in table 2. Generally speaking, ML estimates are slightly smaller than those of equation (4).

Predicting Critical Amino Acid Residues

Our results (see tables 1 and 2) have provided strong statistical evidence for functional divergence after gene duplication (i.e., $\theta_\lambda > 0$). Therefore, it is of great interest to statistically predict which sites are likely to be responsible for these type I functional differences. Indeed, these sites can be further tested by using molecular, biochemical, or transgenic approaches. We shall develop a site-specific profile for this purpose, which can be achieved with a hidden Markov model (HMM), which has been widely used in computational biology from gene finding to pattern recognition (Durbin et al. 1998). Remember that in the two-state model, each site has two possible states, F_0 (functional constraint) and F_1 (functional divergence), with the (prior) probabilities $P(F_1) = \theta_\lambda$ and $P(F_0) = 1 - \theta_\lambda$, respectively. To provide a statistical basis for predicting which state is more likely at a given site, we need to compute the (posterior) probability of state F_1 at this site with X_1 (and X_2) changes in cluster 1 (and 2), $P(F_1 | X_1, X_2)$. Obviously, $P(F_0 | X_1, X_2) = 1 - P(F_1 | X_1, X_2)$. According to the Bayesian law and equations (11) and (14), we can show

$$\begin{aligned} P(F_1 | X_1, X_2) &= \frac{P(F_1)P(X_1, X_2 | F_1)}{P(X_1, X_2)} \\ &= \frac{\theta_\lambda Q_1 Q_2}{(1 - \theta_\lambda)K_{12} + \theta_\lambda Q_1 Q_2}. \end{aligned} \tag{19}$$

Then, given $X_1 = i$ and $X_2 = j$, the posterior (probability) ratio can be defined as follows:

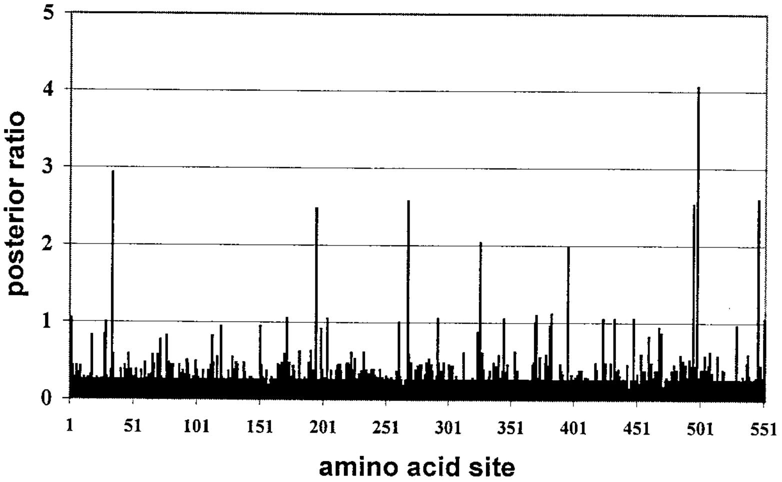

FIG. 5.—The site-specific profile for predicting critical amino acid sites responsible for the (type I) functional divergence between TF and LTF, measured by the posterior probability ratio.

$$R_{ij} = \frac{P(F_1 | X_1 = i, X_2 = j)}{P(F_0 | X_1 = i, X_2 = j)} = \frac{\theta_\lambda}{1 - \theta_\lambda} \frac{Q_1(i)Q_2(j)}{K_{12}(i, j)}, \quad (20)$$

which turns out to be

$$R_{ij} = \frac{\theta_\lambda}{1 - \theta_\lambda} \frac{\Gamma(i + \alpha)\Gamma(j + \alpha)}{\Gamma(i + j + \alpha)} \left(1 + \frac{D_2}{D_1 + \alpha}\right)^i \times \left(1 + \frac{D_1}{D_2 + \alpha}\right)^j \left(1 - \frac{D_1 D_2}{(D_1 + \alpha)(D_2 + \alpha)}\right)^\alpha. \quad (21)$$

We may use either equation (19) or equation (21) to identify these amino acid sites that may be responsible for functional divergence, given a cutoff value. In practice, the choice of a cutoff value is somewhat arbitrary, from $P(F_1 | X_1, X_2) > 0.5$ ($R_{ij} > 1$) to $P(F_1 | X_1, X_2) > 0.95$ ($R_{ij} > 20$). As will be seen below, it may depend on how much information we can obtain.

We have applied equation (21) to the transferrin (TF/LTF) and Myc (N-myc/C-myc) gene families, whose phylogenetic trees are presented in figures 2 and 6, respectively. For example, the site-specific profile R_{ij} for predicting critical amino acid sites responsible for (type I) functional divergence between TF and LTF is plotted against amino acid position (fig. 5). As expected, most amino acid sites have low values, indicating no change in their functional roles after gene duplication. In other words, only a few sites are likely to be involved in functional divergence. Indeed, eight amino acid residues with the highest scores ($R_{ij} > 2$) are apparently considered candidates for type I functional divergence between TF and LTF. Furthermore, after removing these sites, the coefficient of functional divergence is reduced to 0.09, which is not significant from 0. Figure 7 shows the site-specific profile of N-myc/C-myc. The histogram of these posterior ratio values (fig. 8) shows that functional divergence between N-myc/C-myc is affected by only a small number of amino acid residues. Once again, after removing these sites, listed in table 3, the coefficient of functional divergence is reduced to 0.10 (not significant). Indeed, these sites have much more different X_1 and X_2 values, e.g., one of them has no change at all, whereas another one has many changes (table 3), indicating a change in the functional role between the two gene clusters. These predicted residues can be used as targets for further experimentation to determine their functional roles.

Discussion

In this paper, we have studied type I functional divergence after gene duplication, which can be characterized by altered functional constraints between homologous member genes in a gene family. A fundamental measure is the coefficient of functional divergence θ_λ. It can be interpreted as the decrease in rate correlation (r_λ) between two duplicate genes as a result of functional divergence after gene duplication (i.e., $\theta_\lambda = 1 - r_\lambda$), or the (prior) probability that a site is in the F_1 state (functional-divergence-related), i.e., $\theta_\lambda = P(F_1)$. Based on a simple model, we have developed statistical methods for estimating θ_λ and testing whether it is significantly larger than 0, which provides statistical evidence for type I functional divergence. Furthermore, we have developed a site-specific profile (the posterior probability ratio) to predict critical amino acid residues that are responsible for these functional differences by the HMM. Then, given a cutoff value, we can (statistically) identify a group of amino acid residues with the highest scores. Examples involving TF and Myc gene families have shown the potential of our methods in comparative genomics and molecular evolution. For example, these predicted sites can be mapped into the three-dimensional structure of the protein if it is available, and then subsequent biological experimentation can provide a structure-based insight into the underlying mechanism of functional divergence. The current methodology can be directly applied to the study of functional divergence after an evolutionary event such as speciation, domain shuffling, lateral gene transfer, or virus infection, which typically results in a bifurcation in

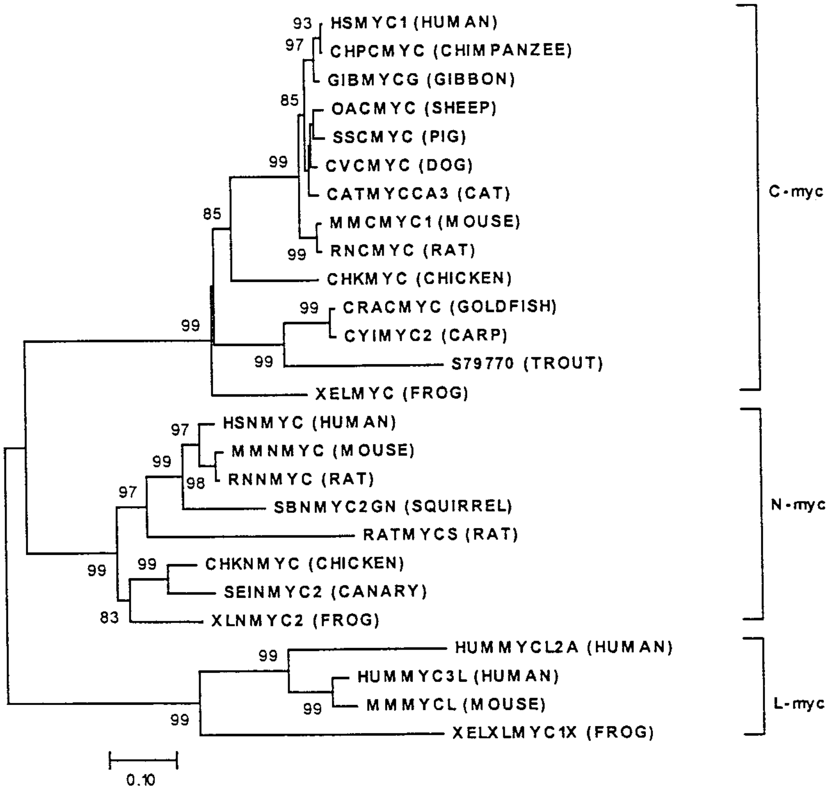

FIG. 6.—The phylogenetic tree of the myc gene family, which was inferred by the neighbor-joining method using amino acid sequences with Poisson distance. Bootstrapping values of >50% are presented.

the phylogenetic tree. For example, the covarion theory of molecular evolution (Fitch and Markowitz 1970) assumes that after speciation, amino acid or nucleotide sites that are invariable in one cluster can be variable in another cluster. However, it is difficult to distinguish between the covarion model and the model of rate variation among sites (Miyamoto and Fitch 1995; Gu and Li 1996). We have recognized that the rate variation among sites provides a background for potential covarion evolution, and the change in functional constraint may occur at a few sites during a particular period of evolution. Since the covarion theory can be treated as a special case of functional divergence, the newly developed method would be helpful in resolving this issue.

When $\theta_\lambda = 0$, the new method is reduced to the gamma distribution model of rate variation among sites. The gamma shape parameter α characterizes the substitution rate variation, which may include the mutation rate variation and the variation in functional constraints among sites (Deng and Fu 1998). Although there are many studies on how to estimate α (e.g., Uzzel and Corbin 1971; Yang 1993; Gu, Fu, and Li 1995; Gu and Zhang 1997), the underlying assumption of no altered functional constraint, which usually does not hold in

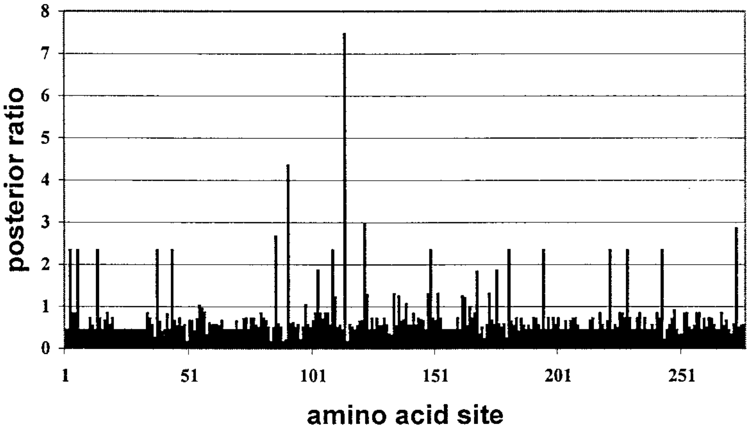

FIG. 7.—The site-specific profile for predicting critical amino acid sites responsible for the (type I) functional divergence between N-myc and C-myc, measured by the posterior probability ratio.

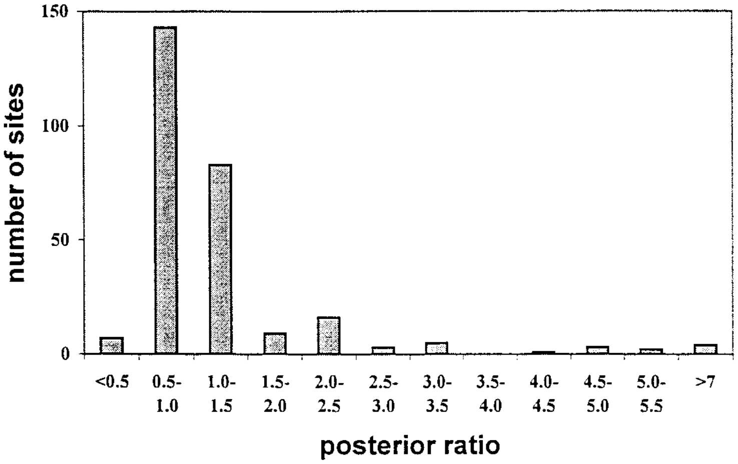

FIG. 8.—The distribution of the posterior ratio over sites in the case of N-myc/C-myc.

gene family evolution (Gu and Li 1998), is not always clearly stated in the literature. In section A.2 of the appendix, we show that when counts from both clusters are simply lumped together, the estimate of parameter α is biased (overestimated) if $\theta_\lambda > 0$ is neglected. Apparently, the current method is more useful in exploring the evolutionary pattern of gene family evolution.

Our theoretical framework can be extended to the Markov chain model (Felsenstein 1981), but it may require a huge amount of computational time. Under the assumption of site-independence, the likelihood function is given by

$$L(\theta_\lambda \,|\, \text{data}) = \prod_k p_k(X), \qquad (22)$$

where $p_k(X)$ is the probability that an amino acid configuration (X) will be observed at site k given the phylogenetic tree. Under the two-state model, $p_k(X)$ can be expressed as

$$p_k(X) = (1 - \theta_\lambda) f_k(X_1, X_2 \,|\, F_0) + \theta_\lambda f_k(X_1, X_2 \,|\, F_1), \qquad (23)$$

where $f_k(X_1, X_2 \,|\, F_0)$ and $f_k(X_1, X_2 \,|\, F_1)$ can be computed as in equation (11) except that the Poisson model is replaced by a Markov chain model (e.g., Kishino, Miyata, and Hasegawa 1990). The current method can be simply called the generalized Gu-Zhang (gGZ) (1997) method, which is a good approximation of the Markov chain model but computationally very fast. Our preliminary result has shown that the MLE of θ_λ based on the Markov chain model is fairly close to that based on the gGZ method (unpublished data). Actually, this is not very surprising, because computer simulation has already shown that the performance of Gu and Zhang's (1997) method (for $\theta_\lambda = 0$) is almost as good as that of the standard Markov chain approach (Gu and Zhang 1997). Indeed, our preliminary result from computer simulation indicates that the current method is asymptotically unbiased and usually is robust if the phylogenetic tree is not accurate (unpublished data). Further study will be focused on how to take the pattern of amino acid changes into account (e.g., Gu, Hewett-Emmett, and Li 1998) and how to include insertions and deletions (indels) in our model (Gu and Li 1995).

Table 3
Amino Acid Sites with the Highest Posterior Ratio Values ($R_{ij} > 2.5$ in Fig. 7) for Type I Functional Divergence Between C-myc and N-myc Genes

	Position	X_1	X_2	R_{ij}
1	253	7.5	0	23.6
2	245	7.0	0	18.5
3	50	0	7.0	14.0
4	176	5.1	0	7.8
5	179	0	4.6	5.1
6	95	0	4.6	5.1
7	244	0	4.5	5.0
8	149	0	4.4	4.8
9	243	3.9	0	4.6
10	118	3.7	0	4.1
11	48	0	3.6	3.5
12	247	7.6	1.3	3.2
13	97	0	3.4	3.2
14	37	0	3.4	3.1
15	56	1.1	7.5	3.1
16	135	7.2	1.3	2.8
17	89	1.1	7.0	2.7

NOTE.—X_1 and X_2 are the expected numbers of changes at a site in gene clusters 1 (C-myc) and 2 (N-myc), respectively, obtained by Gu and Zhang's (1997) method. R_{ij} is the posterior (probability) ratio for the functional divergence at a site with $X_i = i$ and $X_j = j$.

In conclusion, the methodology we have developed in this paper provides a novel approach to exploring the pattern of functional divergence after gene duplication and/or speciation. Moreover, predictions based on the HMM approach can be powerful and very cost-effective in defining a group of amino acid residues that are responsible for these functional differences. If these computational approaches are combined with biological information from functional and structural assays, our understanding about the origins of new functions can be significantly improved.

Acknowledgments

This work was supported by start-up funding from Iowa State University (ISU). The author is grateful to members of the Iowa Computational Molecular Biology Laboratory for valuable discussions, and to Yufeng Wang for assistance in data analysis. A computer program is available on request via e-mail or anonymous ftp at the web server http://www.phyba.iastate.edu which will be included in a newly developing computer package, PHYBA (phylogenetic-based analysis).

APPENDIX

A.1. Derivation of Equation (4)

First, we consider the Poisson process at a given site, in which the first and second moments can be expressed as the following conditional expectations:

$$E[X_i \mid \lambda_i] = \lambda_i T_i,$$
$$E[X_i^2 \mid \lambda_i] = \lambda_i T_i + (\lambda_i T_i)^2 \qquad \text{(A.1)}$$

($i = 1, 2$). If there is no gene conversion or recombination between the two homologous genes, amino acid substitutions at a site are independent between two monophyletic gene clusters, and, therefore,

$$E[X_1 X_2 \mid \lambda_1, \lambda_2] = E[X_1 \mid \lambda_1] \times E[X_2 \mid \lambda_2]. \qquad \text{(A.2)}$$

The evolutionary rates (λ_1 and λ_2) are not only correlated, but also different among sites, which, in principle, can be described by a general joint distribution, $\Phi(\lambda_1, \lambda_2)$. To compute the mean and variance over all sites (for each cluster), let $\phi(\lambda_1)$ and $\phi(\lambda_2)$ be the marginal distributions of $\Phi(\lambda_1, \lambda_2)$ which describe the rate variation among sites. By definition, they are given by $\phi(\lambda_1) = \int_0^\infty \Phi(\lambda_1, \lambda_2)\, d\lambda_2$ and $\phi(\lambda_2) = \int_0^\infty \Phi(\lambda_1, \lambda_2)\, d\lambda_1$, respectively. According to the conditional probability theory, one can show that

$$E[X_i] = E[E[X_i \mid \lambda_i]] = \int_0^\infty \lambda_i T_i \phi(\lambda_i)\, d\lambda_i$$
$$= E[\lambda_i] T_i \qquad \text{(A.3)}$$

($i = 1, 2$), where $E[\lambda_i] = \bar{\lambda}_i$ is the mean rate of λ_i. In the same manner, we have

$$E[X_i^2] = E[\lambda_i] T_i + E[\lambda_i^2] T_i^2 \qquad \text{(A.4)}$$

($i = 1, 2$), where $E[\lambda_i^2] = \int_0^\infty \lambda_i^2 \phi(\lambda_i)\, d\lambda_i$. For simplicity, let $D_i = E[X_i]$ and $V_i = E[X_i^2] - (E[X_i])^2$. From equation (A.4), the variance of λ_i, $\mathrm{Var}(\lambda_i) = E[\lambda_i^2] - (E[\lambda_i])^2$, is given by

$$\mathrm{Var}(\lambda_i) = (V_i - D_i)/T_i^2 \qquad \text{(A.4)}$$

($i = 1, 2$). Now consider the covariance between λ_1 and λ_2. From equations (A.1) and (A.2), we have

$$E[X_1 X_2] = T_1 T_2 \int_0^\infty \lambda_1 \lambda_2 \Phi(\lambda_1 \lambda_2)\, d\lambda_1\, d\lambda_2$$
$$= T_1 T_2 E[\lambda_1 \lambda_2], \qquad \text{(A.6)}$$

and therefore the covariance between X_1 and X_2, σ_{12}, is given by

$$\sigma_{12} = T_1 T_2 \mathrm{Cov}(\lambda_1, \lambda_2). \qquad \text{(A.7)}$$

Then, from equations (A.5) and (A.7), one can easily show that the coefficient of rate correlation r_λ defined by equation (1) is given by

$$r_\lambda = \frac{\sigma_{12}}{\sqrt{(V_1 - D_1)(V_2 - D_2)}}, \qquad \text{(A.8)}$$

which directly leads to equation (4). Since $r_\lambda \leq 1$, we have $\sigma_{12} \leq \sqrt{(V_1 - D_1)(V_2 - D_2)}$, which means

$$r_X = \frac{\sigma_{12}}{\sqrt{V_1 V_2}} \leq r_M = \sqrt{(1 - D_1/V_1)(1 - D_2/V_2)}. \qquad \text{(A.9)}$$

A.2. A Short Note on Rate Variation Among Sites

The gamma distribution model for rate variation among sites assumes no altered functional constraints during evolution, i.e., $\theta_\lambda = 0$. Here, we use a simple case to show that the estimation of the shape parameter α may be biased if the assumption of $\theta_\lambda = 0$ is violated.

In the two-cluster case (fig. 1*A*), let $X = X_1 + X_2$ be the (total) number of changes at a site. One can show that X follows a negative binomial distribution if $\theta_\lambda = 0$, i.e., there are no altered functional constraints (e.g., Gu and Zhang 1997). Under this model, the variance of X is given by

$$V = D + \frac{D^2}{\alpha}, \qquad \text{(A.10)}$$

$$\alpha^* = \frac{\alpha}{1 - b\theta_\lambda} \geq \alpha, \qquad \text{(A.12)}$$

where D is the mean of X. In the same manner for each cluster, we have $V_1 = D_1 + D_1^2/\alpha$ and $V_2 = D_2 + D_2^2/\alpha$ On the other hand, we mention that $X = X_1 + X_2$, such that $D = D_1 + D_2$ and $V = V_1 + V_2 + 2\sigma_{12}$. Since $\sigma_{12} = (1 - \theta_\lambda)\sqrt{(V_1 - D_1)(V_2 - D_2)}$ (see eq. A.8), we have

$$V = V_1 + V_2 + 2(1 - \theta_\lambda)\sqrt{(V_1 - D_1)(V_2 - D_2)}. \qquad \text{(A.11)}$$

Therefore, if we define α^* as $D^2/(V - D)$, one can easily show

$$\alpha^* = \frac{\alpha}{1 - b\theta_\lambda} \geq \alpha, \qquad \text{(A.12)}$$

where $b = 2D_1D_2/(D_1 + D_2)^2$; $\alpha^* = \alpha$ only when $\theta_\lambda = 0$. If we use the method of moments to estimate α under the assumption of no altered functional constraints between these two gene clusters, we obtain $\hat{\alpha} = \hat{D}^2/(\hat{V} - \hat{D})$.According to equation (A.12), for a sufficiently large number of sites, the following relation holds:

$$E[\hat{\alpha}] \approx \alpha^* \geq \alpha, \qquad \text{(A.13)}$$

that is, the estimate of α is biased.

A.3. Time-Dependence of r_X

From equations (6) and (A.5), one can verify that

$$r_X = \frac{r_\lambda}{\sqrt{(1 + a_1/T_1)(1 + a_2/T_2)}}, \qquad \text{(A.14)}$$

where $a_1 = \bar{\lambda}_1/\mathrm{Var}(\lambda_1)$ and $a_2 = \bar{\lambda}_2/\mathrm{Var}(\lambda_2)$.

LITERATURE CITED

Bork, P., and E. V. Koonin. 1998. Predicting functions from protein sequences—where are the bottlenecks? Nat. Genet. **18**:313–318.

Deng, H. W., and Y. X. Fu. 1998. Parsimonious counting of mutations and estimating mutation rate heterogeneity within a DNA sequence. Pp. 64–71 *in* M. K. Uyenoyama and A. von Haeseler, eds. Proceedings of the Trinational Workshop on Molecular Evolution. Duke University Publications Group, Durham, N.C.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, England.

Eisen, J. A. 1998. Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. **8**:163–167.

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

Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

Fitch, W. M., and E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4**:579–593.

Fryxell, K. J. 1996. The coevolution of gene family trees. Trends Genet. **12**:364–369.

Golding, G. B., and A. M. Dean. 1998. The structural basis of molecular adaptation. Mol. Biol. Evol. **15**:355–369.

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., D. Hewett-Emmett, and W. H. Li. 1998. Direct mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica **102/103**: 383–391.

Gu, X., and W. H. Li. 1992. Higher rates of amino acid substitution in rodents than in humans. Mol. Phylogenet. Evol. **1**:211–214.

———. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J. Mol. Evol. **40**:464–473.

———. 1996. A general additive distance with time-reversibility and rate variation among nucleotide sites. Proc. Natl. Acad. Sci. USA **93**:4671–4676.

———. 1998. Evolutionary distances under stationary and nonstationary models of nucleotide substitutions. Proc. Natl. Acad. Sci. USA **95**:5899–5905.

Gu, X., and J. Zhang. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol. Biol. Evol. **14**:1106–1113.

Henikoff, S., E. A. Greene, S. Pietrokovski, P. Bork, T. K. Attwood, and L. Hood. 1997. Gene families: the taxonomy of protein paralogs and chimeras. Science **278**:609–614.

Holland, P. W. H., J. Garcia-Fernandez, N. A. Williams, and A. Sidow. 1994. Gene duplication and the origins of vertebrate development. Dev. Suppl. **1994**:125–133.

Hughes, A. L. 1994. The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. B Biol. Sci. **256**:119–124.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

Kimura, M., and T. Ohta. 1974. On some principles governing molecular evolution. Proc. Natl. Acad. Sci. USA **71**: 2848–2852.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

Li, W. H. 1983. Evolution of duplicated genes. Pp. 14–37 *in* M. Nei and R. K. Koehn, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

———. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

Lundin, L. G. 1993 Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics **16**:1–19.

Miyamoto, M. M., and W. M. Fitch. 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. **12**:503–513.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

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

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, Berlin.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. 1992. Numerical recipes in C. Cambridge University Press, Cambridge, England.

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

Spring, J. 1997. Vertebrate evolution by interspecific hybridisation—are we polyploid? FEBS Lett. **400**:2–8.

Uzzel, T., and K. W. Corbin. 1971. Fitting discrete probability distribution to evolutionary events. Science **172**:1089–1096.

Wakeley, J. 1993 Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37**:613–623.

Wu, C. I., and W. H. Li. 1985. Evidence for higher rates of nucleotide substitution in rodents than in human. Proc. Natl. Acad. Sci. USA **82**:1741–1745.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

Zhang J., H. F. Rosenberg, and M. Nei. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95**:3708–3713.

Yun-Xin Fu, reviewing editor

Accepted August 18, 1999