JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 3, 2001
Mary Ann Liebert, Inc.
Pp. 221–234

Mathematical Modeling for Functional Divergence after Gene Duplication

XUN GU

ABSTRACT

In this paper, I present a statistical framework for modeling the functional divergence after gene duplication. A rate-component model to describe the rate covariation among homologous genes of a gene family is implemented when a phylogenetic tree is known. The Markov chain model is rigorous but may require a huge amount of computational time when the number of sequences is large. On the other hand, the Poisson-based model is mathematically analytical so that computation is very fast even for a large dataset. Moreover, under the posterior framework, we have developed a site-specific profile for predicting important amino acid residues responsible for these functional differences between member genes of a gene family. Our study may have great potential for functional genomics because it is cost-effective, and these predictions can be further tested by biological experimentation.

Key words: functional divergence, gene duplication, Markov chain model, Poisson-based model, posterior prediction.

INTRODUCTION

EXPLORING FUNCTIONAL DIVERGENCE of a gene family after gene duplication is important for postgenomics study (Henikoff *et al.*, 1997; Bork and Koonin, 1998). Indeed, many organisms have undergone genome-wide or local duplication events during their evolution (Ohno, 1970; Lundin, 1993; Holland *et al.*, 1994; Spring, 1997). As a consequence of these gene/genome duplication 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; Henikoff *et al.*, 1997).

It is widely accepted that after gene duplication one gene copy maintains the original function, while the other copy is free to accumulate amino acid changes in the coding region that may lead to functional divergence (Ohno, 1970; Li, 1983). However, how to identify these important sites from sequence analysis remains a challenging problem. The difficulty stems from the fact that most amino acid changes are selectively neutral and are not related to functional divergence (Kimura, 1983; Golding and Dean, 1998). Current bioinformatic tools such as homologous search and phylogenetic reconstruction may not be sufficient to solve this problem (Gu, 1999).

Department of Zoology/Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011.

To detect amino acid changes that are important for functional divergence after gene duplication, we must develop a novel probabilistic model to distinguish between the functional divergence-related change and the background change which mainly represents the neutral evolution. The principle for modeling functional divergence after gene duplication is simple: functional change is highly correlated with the change of evolutionary rate occurring in a certain period of evolution (Gu, 1999). Actually, it is a generalized version of the fundamental rule in molecular evolution and bioinformatics—functional importance is highly correlated with evolutionary conservation (Kimura, 1983). Gu (1999, 2001) proposed statistical methods based on a two-state model for functional divergence. In this paper, we develop an alternative framework (the rate-component model) for modeling the functional divergence after gene duplication. Two probabilistic models for sequence evolution, the Markov chain model and the Poisson-based model, are studied. Some analytical results under the Poisson-based model are obtained and these results facilitate greatly the development of a fast algorithm for large-scale sequence analysis. Furthermore, a site-specific profile based on the posterior prediction is explored and this profile can be very useful for predicting critical amino acid residues for these functional differences between members of a gene family.

FROM DNA SEQUENCE TO FUNCTIONAL DIVERGENCE

Evolutionary perspective of functional divergence

One major issue in functional genomics is how to identify important residues for those functional differences among homologous genes. A cost-effective approach may consist of two steps: (1 a computational sequence analysis to score each residue according to its likelihood to be functional divergence-related and (2 experimentation for a selected group of residues with a given cut-off value. To achieve this goal, we need to develop a novel statistical model for DNA sequence evolution of a gene family so that we can detect the signal from the background of neutral evolution.

We have recognized that *functional divergence may result in the change of the relative importance of residues between homologous genes, which is highly correlated with altered evolutionary rate in DNA sequence evolution.* This is called type I functional divergence. Thus, rate change at an aligned site between two homologous genes can be used as a predictor for functional divergence. Consider a simple case of two gene clusters (Fig. 1). After gene duplication, we assume that the number of changes at a site in each

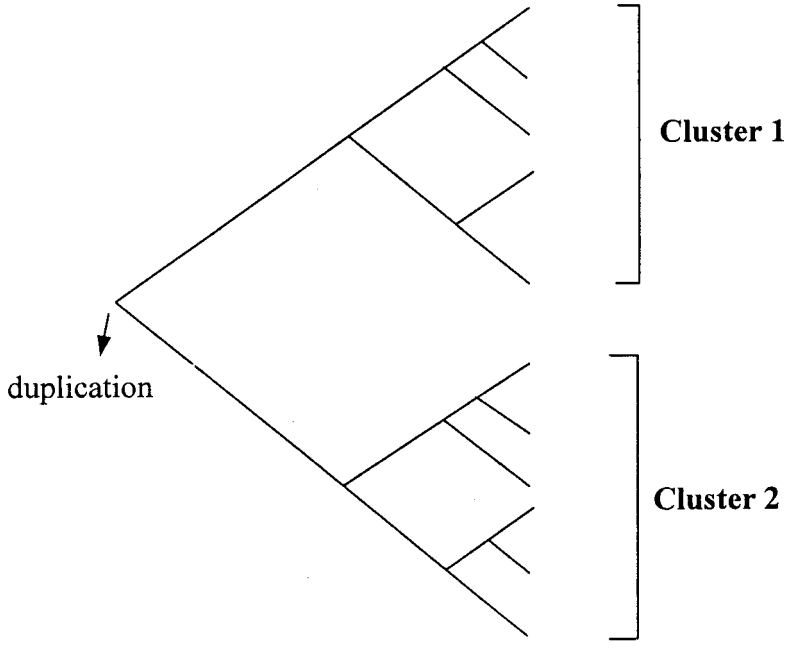

FIG. 1. In the case of two gene clusters after gene duplication.

cluster follows an independent Poisson process, but the Poisson rate is the same between two clusters. Thus, at a given site, the observed difference of the numbers of changes between clusters is mainly caused by two independent random processes. If there are many sites showing very different patterns (e.g., very few changes in one cluster but many in the other) which cannot be explained only by the random process, one may indicate that the evolutionary rates at these sites differ between two gene clusters. Since the evolutionary rate at a site cannot be observed from the sequence data, a statistical model is needed.

Rate-component model for functional divergence

We consider a gene family that has two (paralogous) member genes; each of them has several orthologous sequences that form a monophyletic cluster in the phylogenetic tree (Fig. 1). It is well-known that the evolutionary rate varies among sites due to different functional constraints (Uzzel and Corbin, 1971; Gu *et al.*, 1995; Felsenstein and Churchill, 1996). Moreover, the evolutionary rate at a given site may differ between two gene clusters as a result of functional divergence after gene duplication (Gu, 1999). Our purpose is to model the rate variation among sites and the covariation between homologous gene clusters simultaneously. Let λ_1 and λ_2 be the evolutionary rates of clusters 1 and 2, respectively. We assume that λ_1 and λ_2 can be modeled by the following linear equations:

$$\begin{aligned} \lambda_1 &= (u_0 + u_1)/\beta_1, \\ \lambda_2 &= (u_0 + u_2)/\beta_2, \end{aligned} \tag{1}$$

where the rate components u_0, u_1, and u_2 are independent random variables, and β_1, β_2 are constants. In this three-rate-component model, u_0 describes the rate correlation between λ_1 and λ_2, and u_1 (or u_2) describes the rate independence, respectively. The model for rate variation among sites (e.g., Gu *et al.*, 1995) apparently is a special case when $u_1 = 0$ and $u_2 = 0$, i.e., no independent rate components. From Equation (1) it is easy to show that the covariance between λ_1 and λ_2 is proportional to the variance of u_0, i.e.,

$$Cov(\lambda_1, \lambda_2) = Var(u_0)/\beta_1\beta_2. \tag{2}$$

Furthermore, we assume that each rate component follows a standard gamma distribution with density

$$\pi(u_i) = u_i^{\gamma_i - 1} e^{-u_i} / \Gamma(\gamma_i), \qquad i = 0, 1, 2. \tag{3}$$

The mean and variance of u_i are $E[u_i] = \gamma_i$ and $Var(u_i) = \gamma_i$, respectively. Since $Var(\lambda_i) = \alpha_i/\beta_i^2$ $(i = 1, 2)$, where $\alpha_1 = \gamma_0 + \gamma_1$ and $\alpha_2 = \gamma_0 + \gamma_2$, from Equation (2) one can verify that the coefficient of correlation between λ_1 and λ_2 is given by

$$r_{12} = \frac{\gamma_0}{\sqrt{\alpha_1\alpha_2}}. \tag{4}$$

Thus, the coefficient of functional divergence, which is defined by Gu (1999), is given by

$$\theta_{12} = 1 - r_{12} = 1 - \frac{\gamma_0}{\sqrt{\alpha_1\alpha_2}}. \tag{5}$$

From Equations (1) and (3), it is apparent that λ_1 (or λ_2) is gamma distributed, with densities

$$\phi(\lambda_i) = \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)} \lambda_i^{a_i - 1} e^{-\beta_i \lambda_i} \tag{6}$$

$(i = 1, 2)$. Therefore, λ_1 and λ_2 jointly follow a di-gamma distribution, whose density $\phi(\lambda_1, \lambda_2)$ is given by

$$\phi(\lambda_1, \lambda_2) = \frac{\beta_1\beta_2 e^{-(\beta_1\lambda_1 + \beta_2\lambda_2)}}{\Gamma(\gamma_0)\Gamma(\gamma_1)\Gamma(\gamma_2)} \int_0^A u^{\gamma_0 - 1} (\beta_1\lambda_1 - u)^{\lambda_1 - 1} (\beta_2\lambda_2 - u)^{\gamma_2 - 1} e^u du \tag{7}$$

where $A = \min(\beta_1\lambda_1, \beta_2\lambda_2)$ (see the appendix). For instance, if $\gamma_0 = \gamma_1 = \gamma_2 = 1$, we have

$$\phi(\lambda_1, \lambda_2) = \beta_1\beta_2 e^{-\beta_2\lambda_2}(1 - e^{-\beta_1\lambda_1}), \qquad \text{if } \beta_1\lambda_1 \le \beta_2\lambda_2,$$
$$\phi(\lambda_1, \lambda_2) = \beta_1\beta_2 e^{-\beta_1\lambda_1}(1 - e^{-\beta_2\lambda_2}), \qquad \text{if } \beta_1\lambda_1 > \beta_2\lambda_2. \tag{8}$$

In practice, it is usually more convenient to use the joint density of rate components since u_0, u_1, and u_2 are independent, i.e.,

$$\phi(u_0, u_1, u_2) = \pi(u_0)\pi(u_1)\pi(u_2). \tag{9}$$

THE PROBABILISTIC MODEL OF SEQUENCE EVOLUTION

Under the assumption of site-independence, the likelihood function for a set of multiple-aligned amino acid sequences is given by

$$L(\Theta|data) = \prod_i p(X_1^{(i)}, X_2^{(i)}) \tag{10}$$

where $X_j^{(i)}$ is the amino acid configuration observed in gene clusters j ($j = 1, 2$) at site i, $p(X_1^{(i)}, X_2^{(i)})$ is the probability of (X_1, X_2), and Θ is the parameter set. In the following, we will discuss how to implement the rate-component model into two probabilistic models of sequence evolution, i.e., the Markov chain model and the Poisson-based model.

The Markov chain model

Under the Markov chain model (Felseinstein, 1981; Kishino *et al.*, 1990), $p(X_1, X_2)$ can be derived as follows. First, the transition probability matrix for a given time period t can be computed as $\mathbf{P} = e^{\lambda\mathbf{R}t}$, where λ is the evolutionary rate, and the matrix $\mathbf{R}$ represents the pattern of amino acid changes that can be empirically determined by, for example, the model of Dayhoff *et al.* (1978). Second, conditional on λ_1 (or λ_2), the probability of observing X_1 (or X_2) at a site in clusters 1 (or 2), denoted by $f(X_1|\lambda_1)$ (or $f(X_2|\lambda_2)$), can be computed based on the Markov property (Felsenstein, 1981). Third, by taking expectation with the joint density of λ_1 and λ_2, $p(X_1, X_2)$ is computed by

$$p(X_1, X_2) = \int_0^\infty \int_0^\infty f(X_1|\lambda_1) f(X_2|\lambda_2)\phi(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2 \tag{11}$$

which can be further written as

$$\begin{aligned} p(X_1, X_2) &= \int_0^\infty \int_0^\infty \int_0^\infty f(X_1|u_0, u_1) f(X_2|u_0, u_2)\pi(u_0)\pi(u_1)\pi(u_2) du_0 du_1 du_2 \\ &= E_u[f(X_1|u_0, u_1) f(X_2|u_0, u_2)] \end{aligned} \tag{12}$$

Gu (2001) has studied extensively the Markov chain model for multiple gene clusters, under the two-state model. In principle, the rate-component model can be implemented in a similar way. However, computation of these expectations is rather complicated. Since its applications in sequence analysis are limited when the number of sequences is large, it is desirable to develop a fast algorithm under a simple model (see below).

The Poisson-based model

For each gene cluster, we assume that the number of changes (k) at a given site follows a Poisson process:

$$p_i(k|\lambda_i) = \frac{(\lambda_i T_i)^k}{k!} e^{-\lambda_i T_i}, \qquad k = 0, \ldots, \infty \tag{13}$$

where T_i is the total evolutionary time of gene cluster i ($i = 1, 2$). Note that mathematical treatment for the Poisson-based model is similar to the above model except that the amino acid configuration is reduced to the numer of changes, i.e., $X_1 = k_1$ and $X_2 = k_2$.

The joint distribution. Under the (three) rate-component model, the joint distribution for the numbers of changes ($X_1 = k_1$ and $X_2 = k_2$) is given by

$$\begin{aligned} q(k_1, k_2) &= \int_0^\infty \int_0^\infty p_1(k_1|\lambda_1) p_2(k_2|\lambda_2) \phi(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2 \\ &= E_u[p_1(k_1|\lambda_1) p_2(k_2|\lambda_2)]. \end{aligned} \tag{14}$$

Let ρ_1, ρ_2, and ρ_0 be, respectively

$$\begin{aligned} \rho_1 &= 1 + T_1/\beta_1 = 1 + D_1/\alpha_1 \\ \rho_2 &= 1 + T_2/\beta_2 = 1 + D_2/\alpha_2 \\ \rho_0 &= 1 + T_1/\beta_1 + T_2/\beta_2 = 1 + D_1/\alpha_1 + D_2/\alpha_2 \end{aligned} \tag{15}$$

where D_1 (or D_2) is the average number of changes over sites in gene cluster 1 (or 2). As shown in the appendix, $q(k_1, k_2)$ can be expressed as

$$q(k_1, k_2) = \frac{(\rho_1 - 1)^{k_1} (\rho_2 - 1)^{k_2}}{p_1^{k_1+\gamma_1} \rho_2^{k_2+\gamma_2} \rho_0^{\gamma_0}} G(k_1, k_2) \tag{16}$$

where $G(k_1, k_2)$ is given by

$$G(k_1, k_2) = \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} c_{ij} \left(\frac{\rho_1}{\rho_0}\right)^i \left(\frac{\rho_2}{\rho_0}\right)^j \tag{17}$$

and c_{ij} is given by

$$c_{ij} = \frac{\Gamma(i + j + \gamma_0)\Gamma(k_1 - i + \gamma_1)\Gamma(k_2 - j + \gamma_2)}{i!j!(k_1 - i)!(k_2 - j)!\Gamma(\gamma_0)\Gamma(\gamma_1)\Gamma(\gamma_2)}. \tag{18}$$

Expected number of changes. To apply the Poisson-based model, we need to know the number of changes at each site in each gene cluster (i.e., $X_1 = k_1$ and $X_2 = k_2$). Since k_1 and k_2 cannot be directly observed from the sequence data, a conventional solution is to use the number of minimum-required changes (m) as an approximation, which can be inferred by parsimony under a known phylogenetic tree (Fitch, 1971). However, m is a low-bound of 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 ($\hat{X}$) at a given site is the nonnegative solution of the likelihood equation

$$\sum_{i=1}^{M} \frac{\delta_i b_i}{1 - e^{-\hat{X} b_i / B}} = 1 \tag{19}$$

where B is the total branch length of the gene cluster, b_i is the i-th 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 i-th branch, otherwise $\delta_i = 0$. Extensive computer simulation has shown that the estimate of the 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 (Gu and Zhang, 1997). Two interesting special solutions to Equation (19) are (1) $\hat{X} \approx m$ for short branch lengths, and (2) $\hat{X} = -M \ln(1 - m/M)$ for equal branch lengths.

Method of moments for estimation. When the number of changes at a site in each gene cluster is inferred by Gu and Zhang's (1997) method, say, $X_1 = k_1$ and $X_2 = k_2$, their means, variances, and the covariance can be easily computed, and these are denoted by D_i, $V_i (i = 1, 2)$, and σ_{12}, respectively. Since λ_1 and λ_2 are gamma distributed, it is well known that the shape parameters α_1 and α_2 can be estimated by

$$\hat{\alpha}_i = \frac{D_i^2}{V_i - D_i} \tag{20}$$

$(i = 1, 2)$. Using a similar approach to Gu (1999), one can show that the covariance between $X_1 = k_1$ and $X_2 = k_2$ is given by

$$\sigma_{12} = r_{12} \frac{D_1 D_2}{\sqrt{\alpha_1 \alpha_2}}, \tag{21}$$

from which we can estimate the coefficient of rate correlation (r_{12}) as

$$\hat{r}_{12} = \frac{\sigma_{12}\sqrt{\hat{\alpha}_1 \hat{\alpha}_2}}{D_1 D_2} = \frac{\sigma_{12}}{\sqrt{(V_1 - D_1)(V_2 - D_2)}}. \tag{22}$$

Since $\alpha_1 = \gamma_0 + \gamma_1$, $\alpha_2 = \gamma_0 + \gamma_2$ and $r_{12} = \gamma_0/\sqrt{\alpha_1 \alpha_2}$, from Equations (20) to (22) we have

$$\begin{aligned}
\hat{\gamma}_0 &= \left(\frac{D_1}{V_1 - D_1}\right)\left(\frac{D_2}{V_2 - D_2}\right)\sigma_{12} \\
\hat{\gamma}_1 &= \left(\frac{D_1}{V_1 - D_1}\right)\left(D_1 - \frac{D_2}{V_2 - D_2}\sigma_{12}\right) \\
\hat{\gamma}_2 &= \left(\frac{D_2}{V_2 - D_2}\right)\left(D_2 - \frac{D_2}{V_2 - D_2}\sigma_{12}\right).
\end{aligned} \tag{23}$$

Some examples are shown in Table 1. Each case contains a pair of homologous genes, which are generated in the early stages of vertebrate lineage. Amino acid sequences from vertebrates including mammals, birds, frogs, and fishes are obtained from Genebank. The computer program CLUSTALX is used for the multiple alignment. Based on the phylogenetic tree inferred by the neighbor-joining method

Table 1. Some Examples for Functional Divergence After Gene Duplication[a]

Genes	θ_{12}	α_1	α_2	γ_0	γ_1	γ_2
BMP2/BMP4	0.30	0.60	0.56	0.40	0.20	0.16
CTSK/CTSL	0.37	0.74	0.92	0.52	0.22	0.44
ERβ/ER	0.37	0.43	0.61	0.33	0.10	0.28
IGFII/INS	0.43	0.64	0.82	0.41	0.23	0.41
MDR1/MDR3	0.39	0.50	0.37	0.25	0.25	0.12
MSX2/MSX1	0.40	0.20	0.62	0.21	0.00	0.41
MyoD/MyF5	0.47	0.56	0.66	0.32	0.24	0.33
NF-M/NF-H	0.31	1.42	2.02	1.29	0.13	0.73
EGR2/EGR1	0.34	0.61	0.78	0.37	0.24	0.41
PITX2/PITX1	0.43	0.16	0.29	0.12	0.04	0.17

[a]BMP2/BMP4 are TGFβ-like growth factors; CTSK/CTSL (cathepsin cysteins proteases; ERβ/ER are estrogen receptors; IGFII/INS are insulin-like growth factors; MDR1/MDR3 are multidrug resistance genes; MSX2/MSX1 are homeobox transcription factors; MyoD/MYF5 are bHLH transcription factors; NF-M/NF-H are neurofilaments; EGR2/EGR1 are Zinc finger transcription factors; and PITX2/PITX1 are homeobox transcription factors.

(Saitou and Nei, 1987), the expected number of changes at each site for each gene cluster can be estimated by the method of Gu and Zhang (1997) [also see Equation (19)]. For each pair of homologous genes, we have shown that the coefficient of functional divergence (θ_{12}) is significantly larger than 0 (at the 5% significance level) by the method of Gu (1999). As the shape parameter of rate (α_1 or α_2) typically ranges from 0.16 to 2.02, the range of the shape parameter of the common rate component (γ_0) is from 0.12 to 1.29, and that of the independent rate component (γ_1 or γ_2) is from 0.0 to 0.73. Thus, the relative strength of rate variation among sites and functional divergence can be measured by our method, providing statistical evidence for functional divergence after gene duplication.

Maximum likelihood estimation. The likelihood function under the Poisson-based model can be expressed as follows:

$$L = \prod_i q(k_1^{(i)}, k_2^{(i)}). \tag{24}$$

There are five parameters to estimate, D_1, D_2, γ_0, γ_1, and γ_2. Since no analytical result is possible, a numerical approach is implemented. We can use the estimates of Equation (23) as initial values for the numerical iteration. Since λ_1 and λ_2 are gamma distributed, the marginal distributions $q(k_1)$ and $q(k_2)$ are known to be negative binomial distributed (Gu and Zhang, 1997). Therefore, we can develop a fast algorithm as follows. First, we use Gu and Zhang's (1997) method to obtain the ML estimates for D_i and α_i ($i = 1, 2$). Second, given the relationships $\gamma_1 = \alpha_1 - \gamma_0$ and $\gamma_2 = \alpha_2 - \gamma_0$, a single parameter ($\gamma_0$) can be easily estimated by maximizing Equation (24). Finally, a numerical approach such as the simplex method can be used to obtain the final estimates (Press *et al.*, 1992).

After these ML estimates are obtained, likelihood ratio tests can be constructed to explore the pattern of functional divergence. We consider three special cases that may have biological interest. In case A, the rate component model is reduced to be $\lambda_1 = u_0/\beta_1$ and $\lambda_2 = u_0/\beta_2$ so that $r_{12} = 1$, which represents no functional divergence (or complete rate correlation). In case B, the rate component model is reduced to be $\lambda_1 = u_1/\beta_1$ and $\lambda_2 = u_2/\beta_2$ so that $r_{12} = 0$, which represents no functional correlation (or complete rate independence). And in case C, the rate component model is reduced to be $\lambda_1 = (u_0 + u_1)/\beta_1$ but $\lambda_2 = u_2/\beta_2$, which means functional divergence had occurred exclusively in gene cluster 1. The null hypothesis is $H_0 : \gamma_1 = \gamma_2 = 0$ for case A, $H_0 : \gamma_0 = 0$ for case B, and $H_0 : \gamma_2 = 0$ for case C.

To apply the likelihood ratio test, the joint distribution $q(k_1, k_2)$ under H_0 (cases A to C) needs special treatments because the standard gamma distribution given by Equation (3) cannot be well-defined if the shape parameter tends to be zero. We have solved this problem (see the appendix): for case A, the joint distribution of k_1 and k_2 is given by

$$q(k_1, k_2) = \frac{\Gamma(k_1 + k_2 + \gamma_0)}{k_1!k_2!\Gamma(\gamma_0)} \frac{(\rho_1 - 1)^{k_1}(\rho_2 - 1)^{k_2}}{\rho_0^{k_1+k_2+\gamma_0}}, \tag{25}$$

for case B it is

$$q(k_1, k_2) = \frac{\Gamma(k_1 + \gamma_1)\Gamma(k_2 + \gamma_2)}{k_1!k_2!\Gamma(\gamma_1)\Gamma(\gamma_2)} \frac{(\rho_1 - 1)^{k_1}(\rho_2 - 1)^{k_2}}{\rho_1^{k_1+\gamma_1}\rho_2^{k_2+\gamma_2}}, \tag{26}$$

and for case C it is

$$q(k_1, k_2) = \frac{(\rho_1 - 1)^{k_1}(\rho_2 - 1)^{k_2}}{\rho_1^{k_1+\gamma_1}\rho_2^{k_2+\gamma_2}} \sum_{i=1}^{k_1} b_i \left(\frac{\rho_1}{\rho_0}\right)^i \tag{27}$$

where b_i is given by

$$b_i = \frac{\Gamma(k_2 + i + \gamma_0)\Gamma(k_1 - i + \gamma_1)}{k_2!i!(k_1 - i)!\Gamma(\gamma_1)\Gamma(\gamma_0)}. \tag{28}$$

For instance, we conducted the likelihood ratio test for BMP2/BMP4. The difference of log-likelihood (δ) is $\delta = 8.9$ for case A and $\delta = 13.8$ for case B, indicating that the null hypothesis of no functional divergence, as well as that of complete functional divergence, is statistically rejected. However, we cannot reject the null hypothesis C, i.e., $\gamma_1 = 0.20$ and $\gamma_2 = 0.16$ are virtually the same. The detailed analysis for these ten gene families under the Markov chain model and the Poisson-based model will be published elsewhere.

SITE-SPECIFIC PROFILE: THE POSTERIOR PREDICTION

One major goal for modeling functional divergence after gene duplication is to develop a statistically sound approach for predicting those amino acid residues that are likely to be responsible for these functional differences, which can be further tested by using molecular, biochemical, or transgenic approaches (Gu, 1999). This can be achieved by developing a site-specific profile under the framework of posterior prediction.

To represent the relative importance of an amino acid residue in functional divergence, a straightforward approach is to use the posterior mean of each component. Thus, site-specific profiles based on $E[u_i|X_1, X_2]$ will provide a statistically based score system to define critical amino acid residues for these functional differences.

According to the Bayesian law, the posterior density of each rate component at a site with a given amino acid configuration (X_1, X_2) can be generally expressed as follows:

$$\phi(u_i|X_1, X_2) = \frac{\phi(u_i, X_1, X_2)}{p(X_1, X_2)} = \frac{\pi(u_i)\phi(X_1, X_2|u_i)}{p(X_1, X_2)}, \tag{29}$$

($i = 0, 1, 2$). Under the Poisson-based model, the conditional density for $X_1 = k_1$ and $X_2 = k_2$, i.e., $\phi(k_1, k_2|u_i)$, is given by

$$\begin{aligned}
\phi(k_1, k_2|u_0) &= \int_0^\infty \int_0^\infty p(k_1|u_0, u_1)\, p(k_2|u_0, u_2)\pi(u_1)\pi(u_2)du_1du_2, \\
\phi(k_1, k_2|u_1) &= \int_0^\infty \int_0^\infty p(k_1|u_0, u_1)\, p(k_2|u_0, u_2)\pi(u_0)\pi(u_2)du_0du_2, \text{ and} \\
\phi(k_1, k_2|u_2) &= \int_0^\infty \int_0^\infty p(k_1|u_0, u_1)\, p(k_2|u_0, u_2)\pi(u_0)\pi(u_1)du_0du_1,
\end{aligned} \tag{30}$$

respectively. Similarly to $q(k_1, k_2)$, the analytical form of each conditional density can be obtained (see the appendix for the detailed derivation). To be concise, let

$$F_{ij} = c_{ij}\left(\frac{\rho_1}{\rho_0}\right)^i \left(\frac{\rho_2}{\rho_0}\right)^j / G(k_1, k_2). \tag{31}$$

Since $\sum_{i=0}^{k_1}\sum_{j=0}^{k_2} F_{ij} = 1$, F_{ij} can be interpreted as a posterior spectrum for given k_1 and k_2. Then, the posterior density of u_i conditional of $X_1 = k_1$ and $X_2 = k_2$ is given by

$$\begin{aligned}
\phi(u_0|k_1, k_2) &= \psi(u_0)\sum_{i=0}^{k_1}\sum_{j=0}^{k_2} a_{ij}^{(0)}[\rho_0 u_0]^{i+j}, \\
\phi(u_1|k_1, k_2) &= \psi(u_1)\sum_{i=0}^{k_1} a_i^{(1)}[\rho_1 u_1]^{k_1-i}, \text{ and} \\
\phi(u_2|k_1, k_2) &= \psi(u_2)\sum_{j=0}^{k_2} a_j^{(2)}[\rho_2 u_2]^{k_2-j},
\end{aligned} \tag{32}$$

respectively, where $\psi(u_i)$ $(i = 0, 1, 2)$ is a gamma distribution density with the shape parameter γ_i and the scale parameter ρ_i, i.e.,

$$\psi(u_i) = \frac{\rho_i^{\gamma_i}}{\Gamma(\gamma_i)} u_i^{\gamma_i - 1} e^{-\rho_i u_i} \tag{33}$$

and the constants $a_{ij}^{(0)}$, $a_i^{(1)}$ and $a_j^{(2)}$ are

$$\begin{aligned} a_{ij}^{(0)} &= \frac{\Gamma(\gamma_0)}{\Gamma(i + j + \gamma_0)} F_{ij}, \\ a_i^{(1)} &= \frac{\Gamma(\gamma_1)}{\Gamma(k_1 - i + \gamma_1)} \sum_{j=0}^{k_2} F_{ij}, \text{ and} \\ a_j^{(2)} &= \frac{\Gamma(\gamma_2)}{\Gamma(k_2 - j + \gamma_2)} \sum_{i=0}^{k_1} F_{ij}. \end{aligned} \tag{34}$$

Then, although mathematically tedious, we have shown that the posterior means of rate components,

$$E[u_i | k_1, k_2] = \int_0^\infty u_i \phi(u_i | k_1, k_2) du_i = \frac{\int_0^\infty u_i \phi(u_i, k_1, k_2) du_i}{q(k_1, k_2)}, \tag{35}$$

$(i = 0, 1, 2)$, can be analytically expressed as

$$\begin{aligned} E[u_0 | k_1, k_2] &= \gamma_0/\rho_0 + \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} (i + j) F_{ij}/\rho_0, \\ E[u_1 | k_1, k_2] &= (k_1 + \gamma_1)/\rho_1 - \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} i F_{ij}/\rho_1, \text{ and} \\ E[u_2 | k_1, k_2] &= (k_2 + \gamma_2)/\rho_2 - \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} j F_{ij}/\rho_2, \end{aligned} \tag{36}$$

respectively (see the appendix).

DISCUSSION

In this paper, we have developed the rate-component model for the functional divergence after gene duplication, which has been implemented under the Markov chain model and the Poisson-based model for sequence evolution. In particular, analytical results have been obtained under the Poisson-based model such that a fast algorithm for genomic analysis can be developed. Furthermore, a site-specific profile based on the posterior approach is developed for predicting critical amino acid residues that are responsible for these functional differences between member genes of a gene family. Our study provides a new computational tool for understanding the functional divergence of gene families. For example, these predicted sites can be mapped into the 3D structure of the protein if it is available, and then subsequent biological experimentation can provide a structure-based insight for the underlying mechanism.

We should mention that the new method detects those amino acid sites at which the evolutionary rates (or the functional constraint) differ between two duplicate genes. Within each cluster, the observed amino acid substitution(s) can be neutral. If it is the case, the result simply indicates that neutral spaces (i.e., amino acid types that are allowed at these sites) are different between clusters. It should be noted that

our method cannot detect the underlying mechanisms, such as positive selection, for functional divergence after gene duplication.

The pioneering work for modeling functional divergence during sequence evolution is the covarion hypothesis (Fitch and Markowitz, 1970), which allows the transition between invariable and variable states after speciation. Since the traditional covarion model neglected the role of rate variation between sites, it remains disputable (e.g., Tuffley and Steel, 1998). The current study, as well as our recent work (Gu, 1999) provides an approach to solve this long-lasting issue (Gu, unpublished results).

Although the Markov chain model has nice statistical properties for exploring the pattern of sequence evolution, it is computationally not tractable for a large data set to obtain ML estimates and posterior predictions (Gu, 2001). Since the analytical results under the Poisson-based model have been obtained, it is not difficult to develop a fast algorithm for estimating the level of functional divergence and predicting critical amino acid sites. We will improve the statistical power of our method, for example, by taking the different patterns of amino acid changes into account.

At the current stage, we are focused on the problem of three rate components for two gene clusters. In principle it can be generalized to multiple clusters by the model with more rate compounds. However, the problem of over-parameterization may not be nontrivial. A simple three-step algorithm may be useful to solve the problem of statistical identifiability: first, estimate the level of functional divergence for each pair of gene clusters; second, infer a biologically reasonable pattern of functional divergence from these pairwise comparisons; and third, build a likelihood ratio test and site-specific profile for the prediction. Much work is needed.

ACKNOWLEDGMENTS

The author is grateful to all members in the Center for Bioinformatics and Biological Statistics for valuable discussions. This work was supported by the NIH grant RO1 GM62118 to X.G.

APPENDIX

A.1. Derivation of $\phi(\lambda_1, \lambda_2)$ (Equation 7)

Since u_0, u_1, and u_2 are independent, their joint density is simply given by $\phi(u_0, u_1, u_2) = \pi(u_0)\pi(u_1)\pi(u_2)$, which turns out to be

$$\phi(u_0, u_1, u_2) = \frac{e^{-\sum_{j=0}^{2} x_j}}{\prod_{j=0}^{2} \Gamma(\gamma_j)} \prod_{j=0}^{2} x_j^{\gamma_j - 1}. \tag{A.1}$$

Let $y_1 = u_0 + u_1$ and $y_2 = u_0 + u_2$. One can show that the joint density of u_0, y_1, and y_2 is given by

$$\phi(u_0, y_1, y_2) = \frac{e^{u_0 - \sum_{j=1}^{2} y_j}}{\prod_{j=0}^{2} \Gamma(\gamma_j)} u_0^{\gamma_0 - 1} \prod_{j=1}^{2} (y_j - u_0)^{\gamma_j - 1} \tag{A.2}$$

($y_j > x_j > 0$; $j = 1, 2$). Therefore, by integrating out the variable u_0, the joint density of y_1 and y_2 is given by

$$f(y_1, y_2) = \frac{e^{-\sum_{j=1}^{2} y_j}}{\prod_{j=0}^{2} \Gamma(\gamma_j)} \int_0^{\tilde{y}} u_0^{\gamma_0 - 1} e^{u_0} \prod_{j=1}^{2} (y_j - u_0)^{\gamma_j - 1} du_0 \tag{A.3}$$

where $\tilde{y} = \min(y_1, y_2)$. Since $\lambda_1 = y_1/\beta_1$ and $\lambda_2 = y_2/\beta_2$, it is straightforward to obtain Equation (7).

A.2. Derivation of $q(k_1, k_2)$ (Equation 16)

Since $\lambda_1 = (u_0 + u_1)/\beta_1$, $\lambda_2 = (u_0 + u_2)/\beta_2$, and u_0, u_1, and u_2 are independent, we have

$$\begin{aligned} q(k_1, k_2) &= \int_0^\infty \int_0^\infty \int_0^\infty p_1(k_1) p_2(k_2) \pi(u_0) \pi(u_1) \pi(u_2) du_0 du_1 du_2 \\ &= \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} E\{e^{-(u_1+u_0)T_1/\beta_1} e^{-(u_2+u_0)T_2/\beta_2} (u_1 + u_0)^{k_1} (u_2 + u_0)^{k_2}\}. \end{aligned} \tag{A.4}$$

By noting $(u_1 + u_0)^{k_1} = \sum_{i=0}^{k_1} \binom{k_1}{i} u_0^i u_i^{k_1 - i}$ and $(u_2 + u_0)^{k_2} = \sum_{j=0}^{k_2} \binom{k_2}{j} u_0^j u_1^{k_2 - j}$, Equation (A2) turns out to be

$$q(k_1, k_2) = \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} \binom{k_1}{i} \binom{k_2}{j} A_0 A_1 A_2 \tag{A.5}$$

where $A_0 = E[e^{-u_0(T_1/\beta_1 + T_2/\beta_2)} u_0^{i+j}]$, $A_1 = E[e^{-u_1 T_1/\beta_1} u_1^{k_i - i}]$, and $A_2 = E[e^{-u_2 T_2/\beta_2} u_2^{k_2 - j}]$. One can verify that

$$\begin{aligned} A_0 &= \frac{\Gamma(i + j + \gamma_0)}{\Gamma(\gamma_0)} \frac{1}{(1 + T_1/\beta_1 + T_2/\beta_2)^{i+j+\gamma_0}}, \\ A_1 &= \frac{\Gamma(k_1 - i + \gamma_1)}{\Gamma(\gamma_1)} \frac{1}{(1 + T_1/\beta_1)^{k_1 - i + \gamma_1}}, \text{ and} \\ A_2 &= \frac{\Gamma(k_2 - j + \gamma_2)}{\Gamma(\gamma_2)} \frac{1}{(1 + T_2/\beta_2)^{k_2 - j + \gamma_2}}. \end{aligned} \tag{A.6}$$

Then, after some tedious mathematical simplification, we can verify Equation (16).

In the same manner, the joint distribution of k_1 and k_2 under these special cases, i.e., Equations (25) to (27), can be obtained by noting the following equations.

Case A: Complete correlation

$$\begin{aligned} q(k_1, k_2) &= \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} E\{e^{-u_0(T_1/\beta_1 + T_2/\beta_2)} u_0^{k_1 + k_2}\} \\ &= \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} \frac{\Gamma(k_1 + k_2 + \gamma_0)}{\Gamma(\gamma_0)} (1 + T_1/\beta_1 + T_2/\beta_2)^{-(k_1 + k_2 + \gamma_0)} \end{aligned} \tag{A.7}$$

Case B: Complete independence

$$\begin{aligned} q(k_1, k_2) &= \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} E\{e^{-u_1 T_1/\beta_1} u_1^{k_1}\} E\{e^{-u_2 T_2/\beta_2} u_2^{k_2}\} \\ &= \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} \frac{\Gamma(k_1 + \gamma_1)}{\Gamma(\gamma) 1} \frac{\Gamma(k_2 + \gamma_2)}{\Gamma(\gamma_2)} (1 + T_1/\beta_1)^{-(k_1 + \gamma_1)} (1 + T_2/\beta_2)^{-(k_2 + \gamma_2)} \end{aligned} \tag{A.8}$$

Case C: Unequal functional divergence

$$\begin{aligned} q(k_1, k_2) &= \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} E\{e^{-(u_1+u_0)T_1/\beta_1} e^{-(u_0 T_2/\beta_2)} (u_1 + u_0)^{k_1}\} \\ &= \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} \sum_{i=0}^{k_1} \binom{k_1}{i} E[e^{-u_0(T_1/\beta_1 + T_2/\beta_2)} u_0^{k_2 + i}] E[e^{-u_1 T_1/\beta_1} u_1^{k_1 - i}] \end{aligned} \tag{A.9}$$

A.3. Derivation of posterior means and variances

Let $A_1^{(m)}$, $A_2^{(m)}$, and $A_0^{(m)}$ be defined as follows.

$$\begin{aligned}
A_1^{(m)} &= E[e^{-u_1 T_1/\beta_1} u_1^{k_1-i+m}] \\
&= \frac{\Gamma(k_1+m-i+\gamma_1)}{\Gamma(\gamma_1)} \frac{1}{(1+T_1/\beta_1)^{k_1+m-i+\gamma_1}} \\
A_2^{(m)} &= E[e^{-u_2 T_2/\beta_2} u_2^{k_2-j+m}] \\
&= \frac{\Gamma(k_2+m-j+\gamma_2)}{\Gamma(\gamma_2)} \frac{1}{(1+T_2/\beta_2)^{k_2+m-j+\gamma_2}} \\
A_0^{(m)} &= E[e^{-u_0(T_1/\beta_1+T_2/\beta_2} u_0^{i+j+m}] \\
&= \frac{\Gamma(i+j+m+\gamma_0)}{\Gamma(\gamma_0)} \frac{1}{(1+T_1/\beta_1+T_2/\beta_2)^{i+j+m+\gamma_0}}
\end{aligned} \tag{A.10}$$

Apparently, A_0, A_1, and A_2 given by Equation (A.6) are the special cases of $m = 0$.

First we consider a general formula for any m-order posterior moment of rate component u_1. Note that

$$\int_0^\infty u_1^m \phi(u_1, k_1, k_2) du_1 = \int_0^\infty \int_0^\infty \int_0^\infty u_1^m p_1(k_1) p_2(k_2) f(u_0) f(u_1) f(u_2) du_0 du_1 du_2. \tag{A.11}$$

Similarly to the derivation of $q(k_1, k_2)$, we have shown that it turns out to be

$$\int_0^\infty u_1^m \phi(u_1, k_1, k_2) du_1 = \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} \binom{k_1}{i} \binom{k_2}{j} A_0 A_1^{(m)} A_2. \tag{A.12}$$

In the same manner, we have

$$\begin{aligned}
\int_0^\infty u_2^m \phi(u_1, k_1, k_2) du_2 &= \int_0^\infty \int_0^\infty \int_0^\infty u_2^m p_1(k_1) p_2(k_2) f(u_0) f(u_1) f(u_2) du_0 du_1 du_2 \\
&= \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} \binom{k_1}{i} \binom{k_2}{j} A_0 A_1 A_2^{(m)}
\end{aligned} \tag{A.13}$$

and

$$\begin{aligned}
\int_0^\infty u_0^m \phi(u_1, k_1, k_2) du_2 &= \int_0^\infty \int_0^\infty \int_0^\infty u_0^m p_1(k_1) p_2(k_2) f(u_0) f(u_1) f(u_2) du_0 du_1 du_2 \\
&= \frac{T_1^{k_1} T_2^{k_2}}{\beta_1^{k_1} \beta_2^{k_2} k_1! k_2!} \sum_{i=0}^{k_1} \sum_{j=0}^{k_2} \binom{k_1}{i} \binom{k_2}{j} A_0^{(m)} A_1 A_2.
\end{aligned} \tag{A.14}$$

Therefore, the m-th posterior moment of any rate component u_i is given by

$$E[u_i^m | k_1, k_2] = \frac{\int_0^\infty u_i^m \phi(u_i, k_1, k_2) du_i}{q(k_1, k_2)} \tag{A.15}$$

which turns out to be

$$E[u_1^m|k_1,k_2]=\sum_{i=0}^{k_1}\sum_{j=0}^{k_2}\binom{k_1}{i}\binom{k_2}{j}A_0A_1^{(m)}A_2/\sum_{i=0}^{k_1}\sum_{j=0}^{k_2}\binom{k_1}{i}\binom{k_2}{j}A_0A_1A_2$$

$$E[u_2^m|k_1,k_2]=\sum_{i=0}^{k_1}\sum_{j=0}^{k_2}\binom{k_1}{i}\binom{k_2}{j}A_0A_1A_2^{(m)}/\sum_{i=0}^{k_1}\sum_{j=0}^{k_2}\binom{k_1}{i}\binom{k_2}{j}A_0A_1A_2$$

$$E[u_0^m|k_1,k_2]=\sum_{i=0}^{k_1}\sum_{j=0}^{k_2}\binom{k_1}{i}\binom{k_2}{j}A_0^{(m)}A_1A_2/\sum_{i=0}^{k_1}\sum_{j=0}^{k_2}\binom{k_1}{i}\binom{k_2}{j}A_0A_1A_2. \tag{A.16}$$

Then one can show Equation (36) by noticing the following relations:

$$A_1^{(1)}=\frac{k_1-i+\gamma_1}{1+T_1/\beta_1}A_1,$$

$$A_2^{(1)}=\frac{k_2-j+\gamma_2}{1+T_2/\beta_2}A_2, \text{ and}$$

$$A_0^{(1)}=\frac{i+j+\gamma_0}{1+T_1/\beta_1+T_2/\beta_2}A_0. \tag{A.17}$$

A.4. Deviation of the posterior density

First we mention that the joint distribution (density) $\phi(u_i, k_1, k_2)$ is given by

$$\phi(u_0,k_1,k_2)=\int_0^\infty\int_0^\infty p_1(k_1)p_2(k_2)f(u_0)f(u_1)f(u_2)du_1du_2,$$

$$\phi(u_1,k_1,k_2)=\int_0^\infty\int_0^\infty p_1(k_1)p_2(k_2)f(u_0)f(u_1)f(u_2)du_0du_2, \text{ and}$$

$$\phi(u_2,k_1,k_2)=\int_0^\infty\int_0^\infty p_1(k_1)p_2(k_2)f(u_0)f(u_1)f(u_2)du_0du_1, \tag{A.18}$$

respectively. Similarly to the derivation of $q(k_1, k_2)$ or the posterior mean, one can show

$$\phi(u_0,k_1,k_2)=\sum_{i=0}^{k_1}\sum_{j=0}^{k_2}\binom{k_1}{i}\binom{k_2}{j}A_0^*A_1A_2,$$

$$\phi(u_1,k_1,k_2)=\sum_{i=0}^{k_1}\sum_{j=0}^{k_2}\binom{k_1}{i}\binom{k_2}{j}A_0A_1^*A_2, \text{ and}$$

$$\phi(u_2,k_1,k_2)=\sum_{i=0}^{k_1}\sum_{j=0}^{k_2}\binom{k_1}{i}\binom{k_2}{j}A_0A_1A_2^*, \tag{A.19}$$

where $A_0^*=e^{-u_0\rho_0}u_0^{i+j}\pi(u_0)$, $A_1^*=e^{-u_1\rho_1}u_1^{k_1-i}\pi(u_1)$, and $A_2^*=e^{-u_2\rho_2}u_2^{k_2-j}\pi(u_2)$. Therefore, after tedious algebraic simplification, one can verify Equation (32).

REFERENCES

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

Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C. 1978. A model of evolutionary change in proteins, *in Atlas of Protein Sequence Structure*. Vol. 5, Suppl. 3. National Biomedical Research Foundation, Washington, DC.

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.

Felsenstein, J., and Churchill, G. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* 13, 93–104.

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 Markowitz, E. 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.

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

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

Gu, X. 2001. Maximum likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* 189, 453–464.

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

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

Henikoff, S., Green, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., and Hood, L. 1997. Gene families: The taxonomy of protein paralogs and chimeras. *Science* 278(5338), 609–614.

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

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, England.

Kishino, H., Miyata, T., and 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 *in Evolution of Genes and Proteins*, Sinauer Associates, Sunderland, MA.

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.

Nei, M. 1987. *Molecular Evolutionary Genetics*, Columbia University Press, New York.

Ohno, S. 1970. *Evolution by Gene Duplication*, Springer-Verlag, Berlin.

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

Saitou, N., and Nei, M. 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 Letters* 400, 2–8.

Tuffley, C., and Steel, M.A. 1998. Modelling the covarion hypothesis of nucleotide substitution. *Math. Biosci.* 147, 63–91.

Uzzel, T., and Corbin, K.W. 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.

Address correspondence to:
Xun Gu
Department of Zoology/Genetics
Center for Bioinformatics and Biological Statistics
332 Science II Hall
Iowa State University
Ames, IA 50011

E-mail: xgu@iastate.edu