Maximum-Likelihood Approach for Gene Family Evolution Under Functional Divergence

Xun Gu

Department of Zoology/Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University

According to the observed alignment pattern (i.e., amino acid configuration), we studied two basic types of functional divergence of a protein family. Type I functional divergence after gene duplication results in altered functional constraints (i.e., different evolutionary rate) between duplicate genes, whereas type II results in no altered functional constraints but radical change in amino acid property between them (e.g., charge, hydrophobicity, etc.). Two statistical approaches, i.e., the subtree likelihood and the whole-tree likelihood, were developed for estimating the coefficients of (type I or type II) functional divergence. Numerical algorithms for obtaining maximum-likelihood estimates are also provided. Moreover, a posterior-based site-specific profile is implemented to predict critical amino acid residues that are responsible for type I and/or type II functional divergence after gene duplication. We compared the current likelihood with a fast method developed previously by examples; both show similar results. For handling altered functional constraints (type I functional divergence) in the large gene family with many member genes (clusters), which appears to be a normal case in postgenomics, the subtree likelihood provides a solution that is computationally feasible and robust against the uncertainty of the phylogeny. The cost of this feasibility is the approximation when frequencies of amino acids are very skewed. The potential bias and correction are discussed.

Introduction

One major goal in functional genomics is to explore functional divergence of a gene family among member genes (Henikoff et al. 1997; Bork and Koonin 1998). Many organisms, from yeast to humans, have undergone genomewide or local chromosome duplication events during their evolution (Ohno 1970; Lundin 1993; Holland et al. 1994; Spring 1997; Wolfe and Shields 1997). After gene duplication, one gene copy maintains the original function, while the other copy is free to accumulate amino acid changes toward functional divergence (Li 1983). As a result, many genes are represented as several paralogs in the genome with related but distinct functions.

Since gene family proliferation is thought to have provided the raw materials for functional innovations, it is desirable, from sequence analysis, to identify amino acid sites that are responsible for the functional diversity. This approach has great potential for functional genomics because it is cost-effective, and these predictions can be further tested by experimentation. For molecular evolutionists, it is important to know the level of functional divergence after gene (or genome) duplication, as well as how many amino acid substitutions are actually involved in functional innovations. Since most amino acid changes are not related to functional divergence but represent neutral evolution, it is crucial to develop appropriate statistical methods to distinguish between these two possibilities. Indeed, some conventional computational approaches, such as homologous search and multiple alignment, may not be sufficient to solve this problem, because amino acid differences between two duplicate gene copies can be the result of either an ancient gene duplication or a more recent rapid divergence (Golding and Dean 1998).

It has been recognized that when sequences of a gene family are available, the identification of functionally important residues can be approached computationally (e.g., Casari, Sander, and Valencia 1995; Lichtarge, Bourne, and Cohen 1996; Livingstone and Barton 1996; Gu 1999; Landgraf, Fischer, and Eisenberg 1999). The approach introduced by Casari, Sander, and Valencia (1995) used a vectorial analysis of sequence profiles to identify functionally important residues. Lichtarge, Bourne, and Cohen (1996) developed a method called evolutionary tracing, which was extended recently by Landgraf, Fischer, and Eisenberg (1999), known as weighted evolutionary tracing. In these methods, the degree of conservation in each position is scored for different subfamilies and then visualized on the three-dimensional protein structure.

Gu (1999) has developed a novel probabilistic model based on the underlying principle that functional divergence after gene duplication is highly correlated with the change of evolutionary rate. This correlation is the complement to a fundamental rule in molecular evolution—functional importance is highly correlated with evolutionary conservation (Kimura 1983). A site-specific profile based on posterior probability was then developed to predict critical residues for functional differences between two gene clusters. Wang and Gu (unpublished data) have successfully applied this newly developed method to the study of the functional diversity of the caspase gene family and found that our predictions are supported by experimental data.

In this paper, the modeling for functional divergence after gene duplication was studied extensively under the Markov chain model of sequence evolution (Felsenstein 1981), which is further extended to the case of a large family with many member genes. According to the observed alignment pattern (amino acid configuration), we studied two important types of functional divergence (type I and type II, respectively). We show that

Key words: gene family evolution, functional divergence, gene (genome) duplication, maximum likelihood, functional prediction.

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

Mol. Biol. Evol. 18(4):453–464. 2001

	Sequence	Type 0	Type 1	Type 2	Type-U
Gene 1	1	CR	**WQ**LV	RV	KTLI
	2	CR	**WQ**IV	RV	RVLI
	3	CR	**WQ**VG	RV	KIIV
	4	CR	**WQ**VG	RV	NVLL
	5	CR	**WQ**AT	RV	DMLL
	6	CR	**WQ**AT	RV	IKIL
	7	CL	**WQ**VI	RV	EKLI
	8	CR	**WQ**IT	RV	DLVL
Gene 2	9	CR	LT**FD**	DR	LKLM
	10	CR	IT**FD**	DR	QLLV
	11	CR	IT**FD**	ER	RLVV
	12	CR	YS**FD**	DK	LHVV
	13	CR	LE**FD**	DR	KMAL
	14	CL	LE**FE**	DR	KLLI
	15	CR	LE**FD**	DR	KLLL
	16	CR	VG**FD**	DK	ELII
	17	CR	VT**FD**	DR	RLII

FIG. 1.—A hypothetical multiple alignment to show universally conserved sites (type 0), type I and type II amino acid configurations, and type U sites (unclassified).

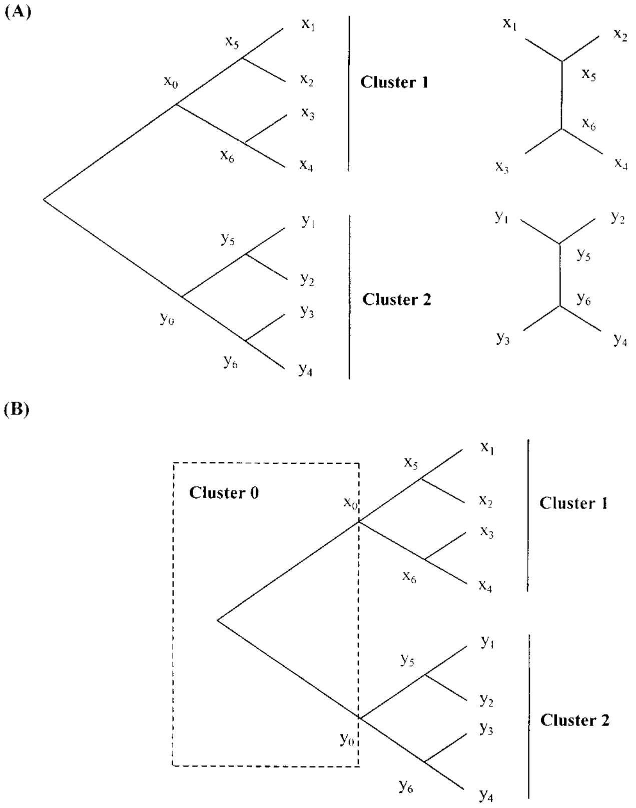

FIG. 2.—Gene clusters for the subtree likelihood (*A*) and for the whole-tree likelihood (*B*). The early stage after gene duplication is designated cluster 0.

Gu's (1999) method is a fast algorithm for two gene clusters. The performance of these methods was compared by examples.

Functional Divergence After Gene Duplication

Consider a multiple alignment of a gene family with two homologous genes, genes 1 and 2 (fig. 1). Although various terminologies have been used (e.g., Casari, Sander, and Valencia 1995; Lichtarge, Bourne, and Cohen 1996; Livingstone and Barton 1996; Gu 1999; Landgraf, Fischer, and Eisenberg 1999), amino acid configurations can be tentatively classified as follows. Type 0 represents amino acid configurations that are universally conserved through the whole gene family, implying that these residues are important for the common function shared by all member genes. Type I represents amino acid configurations that are very conserved in gene 1 but highly variable in gene 2, or vice versa, implying that these residues have experienced altered functional constraints. Type II represents amino acid configurations that are very conserved in both genes but whose biochemical properties are very different, e.g., charge positive versus negative, implying that these residues may be responsible for functional specification. Finally, amino acid configurations at many residues are not so clear-cut, and they have to be regarded as unclassified (type U in fig. 1).

Because of apparent functional implications, Livingstone and Barton (1996) proposed an ad hoc search algorithm to define these types of amino acid configurations. However, there are several potential problems. First, the score system of Livingstone and Barton (1996) is subject to arbitrariness. Second, negligence of the phylogenetic tree may cause bias (false positive). Third, one cannot rule out the possibility that these types are actually generated by a stochastic nature of sequence evolution. As shown below, instead of assigning a definite type to each site, we will establish a probabilistic relationship between types of functional divergence and amino acid configuration.

After gene duplication, functional divergence between homologous genes 1 and 2 is likely to occur in the early stage (fig. 2). According to amino acid configurations that are likely to be generated, there are two basic types of functional divergence after gene duplication. Type I functional divergence results in altered functional constraints (i.e., different evolutionary rates) between duplicate genes. Type II functional divergence results in no altered functional constraints but a radical change in amino acid properties between them (e.g., charge, hydrophobicity, etc.). Intuitively, one may expect that type I (or type II) amino acid configurations are likely to be generated by type I (or type II) functional divergence. This might be true only when the effect of type I (or type II) functional divergence has been shown to be statistically significant under a stochastic model. Thus, the possibility of a site being functional divergence–related (type I or type II) can be measured by a posterior probability when the observed amino acid configuration is given.

Type I Functional Divergence: Two Clusters

Following the statistical framework developed by Gu (1999), we build a ''subtree'' likelihood to estimate the (type I) functional divergence by detecting the level of altered functional constraints between two clusters (fig. 2*A*). The advantage of subtree likelihood is its simplicity, because the phylogenetic relationship among gene clusters will not be considered.

Unrooted Likelihood for Each Cluster

Under the Markov chain model, the likelihood for sequence evolution can be derived as follows (Felsenstein 1981; Kishino, Miyata, and Hasegawa 1990). First, the transition probability matrix for a given time period t can be computed as $\mathbf{P} = e^{\lambda \mathbf{R} t}$, where the rate matrix $\mathbf{R}$ represents the pattern of amino acid substitutions, which can be empirically determined by, for example, the Dayhoff model (Dayhoff, Schwartz, and Orcutt 1978) or the JTT model (Jones, Taylor, and Thornton 1992). The evolutionary rate (λ) may vary among sites because of different functional constraints. Usually, λ is treated as a random variable which follows a gamma distribution, i.e.,

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

(Uzzel and Corbin 1971). The shape parameter α describes the strength of rate variation among sites (i.e., a small value of α means a strong rate heterogeneity among sites, and $\alpha = \infty$ means no rate variation among sites), whereas β is a scale constant (Yang 1993; Gu, Fu, and Li 1995).

Consider the phylogenetic tree in figure 2. Let $X = (x_1, x_2, x_3, x_4)$ and $Y = (y_1, y_2, y_3, y_4)$ be the observed amino acid configurations of a site for clusters 1 and 2, respectively. For the unrooted subtree for cluster 1 or 2 (fig. 2*A*), the conditional probability of observing X or Y at a site can be written as follows:

$$f(X|\lambda) = \sum_{x_5=1}^{20} \sum_{x_6=1}^{20} b_{x_5} P_{x_5 x_1} P_{x_5 x_2} P_{x_5 x_6} P_{x_6 x_3} P_{x_6 x_4}$$

$$f(Y|\lambda) = \sum_{y_5=1}^{20} \sum_{y_6=1}^{20} b_{y_5} P_{y_5 y_1} P_{y_5 y_2} P_{y_5 y_6} P_{y_6 y_3} P_{y_6 y_4}, \tag{2}$$

respectively, where $P_{ij} = P_{ij}(v_{ij})$ is the transition probability from node i to node j, v_{ij} is the branch length between them, and b_i is the frequency of amino acid i. By integrating out the random variable λ, the probability of observing X or Y at a site is given by

$$p(X) = E[f(X|\lambda)] = \int_0^{\infty} f(X|\lambda)\phi(\lambda)\, d\lambda$$

$$p(Y) = E[f(Y|\lambda)] = \int_0^{\infty} f(Y|\lambda)\phi(\lambda)\, d\lambda, \tag{3}$$

respectively, where E means taking expectation.

Two-State Model for Functional Divergence

For two gene clusters generated by gene duplication, the two-state model assumes that in each cluster one site has two possible states, F_0 (functional divergence–unrelated) and F_1 (functional divergence–related). As a result, there are four combined states, i.e., (F_0, F_0), (F_0, F_1), (F_1, F_0), and (F_1, F_1). These states are also called functional divergence configurations, where the first position is for cluster 1 and the second is for cluster 2. When a site is under (F_0, F_0), i.e., no altered functional constraints in both clusters, the evolutionary rate at this site is virtually the same between two clusters, i.e., $\lambda_1 = \lambda_2$. For the last three combined states, however, the amino acid residue has experienced altered functional constraints (i.e., under F_1) at least in one cluster, resulting in statistical independence between λ_1 and λ_2 (Gu 1999).

The assumption of rate-independence for type I functional divergence means that knowledge of the evolutionary rate (or the functional constraint) at such sites in one cluster provides no information for predicting the intensity of functional constraint in the other cluster. Since λ_1 and λ_2 are independent under (F_0, F_1), (F_1, F_0), or (F_1, F_1), these combined states are not distinguishable under the current model; they have to be degenerated to a single one. Consequently, there are two nondegenerate combined states (functional divergence configurations), denoted by $S_0 = (F_0, F_0)$ and $S_1 = (F_0, F_1) \cup (F_1, F_0) \cup (F_1, F_1)$, respectively. It should be noted that (F_0, F_0) was written as $\mathbf{F}_0$, and S_1 was written as $\mathbf{F}_1$ (Gu 1999). To avoid further confusion, the F-notation describes the status in a single cluster, while S-notation is used for the functional divergence configuration of a gene family.

The Subtree Likelihood

Let $P(S_1) = \theta_{12}$ be the probability of a site being in state S_1, and let $P(S_0) = 1 - \theta_{12}$ be the probability of a site being in state S_0. We call θ_{12} the coefficient of type I functional divergence between cluster 1 and cluster 2 (Gu 1999). Let X and Y be the amino acid configurations of a site in clusters 1 and 2, respectively. Our purpose is to build a likelihood function for estimating θ_{12} from sequences. Gu (1999) has shown that the subtree likelihood provides a simple solution for this purpose. Since it only depends on the (unrooted) subtrees of two clusters, the joint probability can be easily derived based on the pattern of rate independence. In the following, we use an asterisk to distinguish the subtree likelihood from the conventional (whole-tree) likelihood, e.g., the joint probability of two subtrees is denoted by $p^*(X, Y)$.

Since evolutionary rates (λ_1 and λ_2) at an S_1 site (i.e., a site under S_1) are statistically independent between two clusters, whereas they are completely correlated ($\lambda_1 = \lambda_2 = \lambda$, without loss of generality) at an S_0 site, the joint probability of subtrees conditional on S_0 or S_1 is given by

$$\begin{aligned} f^*(X, Y|S_0) &= \int_0^{\infty} f(X|\lambda) f(Y|\lambda)\phi(\lambda)\, d\lambda \\ &= E[f(X|\lambda)f(X|\lambda)] \\ f^*(X, Y|S_1) &= p(X)p(Y) \\ &= E[f(X|\lambda_1)] \times E[f(Y|\lambda_2)], \end{aligned} \tag{4}$$

where $f(X|\lambda_1)$ or $f(Y|\lambda_2)$ is the likelihood for each unrooted subtree, respectively; e.g., it is given by equation (2) for the phylogeny in figure 2*A*.

It should be noted that equation (4) actually assumes that the prior frequencies of amino acids at a site are independent between clusters *X* and *Y*, which means that the internal branch length is large. As shown in appendix A, the subtree likelihood (denoted by f^0) can also be built if this prior is identical (i.e., the internal branch length is small). Interestingly, we have shown that the whole-tree likelihood can be expressed as a linear combination of f^* and f^0 (see eq. A.4). Here, we use f^* as an example. Intuitively, it usually gives more conservative results than f^0.

From the two-state model, one can easily show that the joint probability of two subtrees can be written as

$$p^*(X, Y) = (1 - \theta_{12}) f^*(X, Y | S_0) + \theta_{12} f^*(X, Y | S_1). \quad (5)$$

Then, under the assumption of site-independence, the likelihood function over all sites (gaps excluded) is given by

$$L^*(\mathbf{x} | \text{data}) = \prod_k p^*(X^{(k)}, Y^{(k)}), \quad (6)$$

where *k* runs for sites, and **x** is the set of unknown parameters.

Numerical Algorithm

It is complicated to compute $p^*(X, Y)$; it involves the phylogenetic tree, branch lengths (**v**), the shape parameter (α) of a gamma distribution, and the coefficient of functional divergence (θ_{12}). Usually, the integration requires a great deal of computational time even for a small number of sequences (Gu, Fu, and Li 1995). We propose the following iteration to solve this problem:

1. The phylogenetic tree is inferred by the neighbor-joining method (Saitou and Nei 1987), which has a performance similar to that of the maximum-likelihood method but can handle very large numbers of sequences.
2. Given the inferred topology, the branch lengths (**v**) are estimated by a least-squares method, and the gamma shape parameter (α) is estimated by Gu and Zhang's (1997) method. Then, computation of those expectations in equation (4) can be approximated similar to Yang (1994).
3. Regarding all other parameters as constants, the maximum-likelihood estimate (MLE) of θ_{12} can be obtained by $\partial \ln L^*/\partial \ln \theta_{12} = 0$, which satisfies

   $$\sum_{k=1}^{N} \frac{1}{h_k + \hat{\theta}_{12}} = 0 \quad (7)$$

   where *N* is the sequence length and $h_k = 1/(a_k - 1)$, $a_k = f^*(X, Y | S_1)/f^*(X, Y | S_0)$ for site *k*.
4. A numerical iteration, such as the simplex method (Press et al. 1992), is implemented to find the final maximum-likelihood (ML) estimates of **v**, α, and θ_{12} under the given phylogeny.

After obtaining these ML estimates, the likelihood ratio test (LRT) can be constructed under the null hypothesis H_0: $\theta_{12} = 0$ versus H_A: $\theta_{12} > 0$. If H_0 is rejected significantly, it provides statistical evidence for functional divergence in the coding region after gene duplication; i.e., functional constraints have shifted between two homologous genes.

Table 1
Combined States (functional divergence configurations) for the Subtree Likelihood with Three Gene Clusters

State (S_i)	F_0/F_1	$P(S_i)$	Rate independence[a]	Type I Functional Divergence[b]
S_0	(F_0, F_0, F_0)	π_0	$\lambda_1 = \lambda_2 = \lambda_3$	No
S_1	(F_1, F_0, F_0)	π_1	λ_1, $\lambda_2 = \lambda_3$	Cluster 1
S_2	(F_0, F_1, F_0)	π_2	$\lambda_1 = \lambda_2$, λ_3	Cluster 2
S_3	(F_0, F_0, F_1)	π_3	λ_1, $\lambda_2 = \lambda_3$	Cluster 3
S_4	(F_0, F_1, F_1)	π_4	λ_1, λ_2, λ_3	Clusters 2, 3
	(F_1, F_0, F_1)	π_4	λ_1, λ_2, λ_3	Clusters 1, 3
	(F_1, F_1, F_0)	π_4	λ_1, λ_2, λ_3	Clusters 2, 3
	(F_1, F_1, F_1)	π_4	λ_1, λ_2, λ_3	Clusters 1, 2, 3

[a] Rate independence under each state can be illustrated by the following example: λ_1, $\lambda_2 = \lambda_3$ means that λ_1 is independent of λ_2 or λ_3.
[b] Indicates which cluster(s) is under (type I) functional divergence.

Type I Functional Divergence: Multiple Clusters

For a large gene family with many member genes (clusters), the pattern of amino acid configuration is complicated. Even for three clusters, type I amino acid configuration contains many subtypes: cluster 1 is highly variable but clusters 2 and 3 are conserved, etc. The subtree likelihood can be extended to any *n* gene clusters but may demand much computational time when *n* is large.

Subtree Likelihood for Multiple Clusters

Let λ_i and X_i be the evolutionary rate and amino acid configuration of a site in cluster *i*, respectively; $i = 1, \ldots, n$, and $X = (X_1, \ldots, X_n)$. For two possible states (F_0/F_1) in each gene cluster, we have, in total, 2^n possible combined states (functional divergence configurations).

For three gene clusters ($n = 3$), all functional divergence configurations ($2^3 = 8$) are listed as follows: (F_0, F_0, F_0), (F_0, F_0, F_1), (F_0, F_1, F_0), (F_1, F_0, F_0), (F_0, F_1, F_1), (F_1, F_0, F_1), (F_1, F_1, F_0), and (F_1, F_1, F_1), where the first, second, and third positions refer to the F_0/F_1 status of gene clusters 1, 2, and 3, respectively. For each of them, the relationship of evolutionary rates (λ_1, λ_2, and λ_3) among three clusters is shown in table 1. If a site is under (F_0, F_0, F_0), λ_1, λ_2, and λ_3 are completely correlated so that one can assume $\lambda_1 = \lambda_2 = \lambda_3$ ($=\lambda$) without loss of generality. Similarly, under (F_0, F_0, F_1), $\lambda_1 = \lambda_2$ ($=\lambda$), which is independent of λ_3, and so forth. However, the last four combined states, (F_0, F_1, F_1), (F_1, F_0, F_1), (F_1, F_1, F_0), and (F_1, F_1, F_1), have to be degenerated to a single combined state (S_4) because λ_1, λ_2, and λ_3 are mutually independent (table 1). Note that there are $m = 2^n - n$ nondegenerate combined states (functional divergence configurations) in the case of *n* clusters, which are denoted by S_j, $j = 0, \ldots, m - 1$.

In particular, S_0 indicates the functional divergence configuration that all clusters are under F_0, i.e., $S_0 = (F_0, F_0, \ldots, F_0)$.

According to the pattern of rate independence (table 1), we can show that the joint probability of three subtrees under each functional divergence configuration (S_j) is given by

$$\begin{aligned} f^*(X|S_0) &= E[f(X_1|\lambda)f(X_2|\lambda)f(X_3|\lambda)] \\ f^*(X|S_1) &= E[f(X_1|\lambda_1)] \times E[f(X_2|\lambda)f(X_3|\lambda)] \\ f^*(X|S_2) &= E[f(X_2|\lambda_2)] \times E[f(X_1|\lambda)f(X_3|\lambda)] \\ f^*(X|S_3) &= E[f(X_3|\lambda_3)] \times E[f(X_1|\lambda)f(X_2|\lambda)] \\ f^*(X|S_4) &= E[f(X_1|\lambda_1)] \times E[f(X_2|\lambda_2)] \\ &\quad \times E[f(X_3|\lambda_3)], \end{aligned} \tag{8}$$

where $f(X_1|\lambda)$, $f(X_2|\lambda)$, or $f(X_3|\lambda)$ is the likelihood for the (unrooted) subtree of each gene cluster, respectively.

Let π_j be the (prior) probability of a site under S_j, i.e., $\pi_j = P(S_j)$. Thus, the joint probability of three subtrees at a site is given by

$$p^*(X) = \sum_{j=0}^{m-1} \pi_j f^*(X|S_j), \tag{9}$$

where $m = 5$. Apparently, equation (5) is a special case of equation (9) when $n = 2$ (and so $m = 2$), and $\pi_0 = 1 - \theta_{12}$ and $\pi_1 = \theta_{12}$. In general, π_j ($j = 1, 2, \ldots, m - 1$) are called the coefficients of type I functional divergence for functional divergence configuration S_j. In particular, we define

$$\pi_f = 1 - \pi_0 = \sum_{j=1}^{m-1} \pi_j \tag{10}$$

as the coefficient of (type I) functional divergence of the gene family.

Numerical Algorithm

Similar to equation (6), unknown parameters can be estimated by maximizing the likelihood $L^* = \Pi_k p^*(X^{(k)})$, which can be approached by extending the algorithm for two clusters, i.e., equation (7) is replaced by the following procedure. When the number of gene clusters (n) is not very large, the Newton-Raphson algorithm is computationally efficient. Let π be the parameter vector, $\pi = (\pi_1, \ldots, \pi_{m-1})$. The iteration equation is then given by $\pi^{(l+1)} = \pi^{(l)} - \mathbf{H}^{-1}\mathbf{g}$, where $\mathbf{g}$ is the grade vector, whose ith element is $g_i = \partial \ln L^*/\partial\pi_i$, and $\mathbf{H}$ is the Haesson matrix, whose ijth element is $H_{ij} = \partial^2 \ln L^*/\partial\pi_i\partial\pi_j$. When appropriate initial values are given, $\pi^{(l)}$ would converge to $\hat{\pi}$, as $l \to \infty$. Finally, their large sample variance-covariance matrix can be approximately estimated by the inverse of Fisher's information matrix.

When n is large, an expectation and maximization (EM) algorithm can be implemented (Dempster, Laird, and Rubin 1977). The EM method is a very general iterative approach for the data set with missing (or incomplete) data. In our case, the ML estimates of π_i will be easy to obtain if we know the state (F_0/F_1) to which each site belongs in each gene cluster. Thus, the original data set is treated as incomplete data, missing the category information. Using a current estimate of the unknown parameter values, the expected value of the incomplete data is computed, weighted by the posterior probability. This is the expectation [E] step. The result is a set of likelihood equations that are considerably easier to solve than the full likelihood (the maximization [M] step). The new estimates obtained from the M step are then used to update the expected values, and this approach is iterated until convergence. A detailed discussion of implementing an EM algorithm for a large gene family will be published elsewhere.

LRTs can be constructed under various null hypotheses by specifying some coefficients of functional divergence. In particular, the LRT under the null H_0: $\pi_0 = 1$ (or $\pi_f = 0$) is apparently the most powerful test. Significant rejection of the null hypothesis suggests functional divergence among member genes of a gene family.

Type I and II Functional Divergences: Two Gene Clusters

In spite of the efficiency in estimating type I functional divergence, the subtree likelihood is not able to detect type II functional divergence that requires the evolutionary relationship between clusters. Therefore, it is desirable to build a "whole-tree" likelihood for estimating these two types of functional divergence simultaneously.

In the early stage after gene duplication, many evolutionary forces (e.g., positive selection, functional relaxation, or coevolution between contact sites) might play roles in amino acid substitutions such that a comprehensive modeling could be complicated. A simple solution is to consider the internal branch between two clusters (i.e., the early stage) as cluster 0 (fig. 2*B*), which is ancestral. Let λ_1 and λ_2 be the evolutionary rates in clusters 1 and 2, respectively, and let λ_0 be the evolutionary rate in the internal branch (cluster 0). For each cluster, a given site has two possible states, F_0 (functional divergence–unrelated) and F_1 (functional divergence–related). Therefore, similar to the subtree likelihood of three clusters, we have $2^3 = 8$ possible combined states that can be degenerated to 5 functional divergence configurations, under which the relationship between λ_0, λ_1, and λ_2 is shown in table 2. Let π_j ($j = 0, 1, \ldots, 4$) be the probability of a site being under S_j, i.e., $\pi_j = P(S_j)$. For a given site, the conditional probability of observing X and Y is given by

$$f(X, Y|\lambda) = \sum_{x_0=1}^{20} \sum_{y_0=1}^{20} b_{x_0} P_{x_0 y_0}(\upsilon|\lambda_0) f(X|\lambda_1; x_0) f(Y|\lambda_2; y_0), \tag{11}$$

where $f(X|\lambda_1; x_0)$ and $f(Y|\lambda_2; y_0)$ are the likelihood functions for clusters 1 and 2, conditional on the roots x_0

Table 2
Combined States (functional divergence configurations) for the Whole-Tree Likelihood with Two Gene Clusters

State (S_i)	F_0/F_1	$P(S_i)$	Rate independence[a]	Functional Divergence[b]
S_0	(F_0, F_0, F_0)	π_0	$\lambda_1 = \lambda_2 = \lambda_3$	No
S_1	(F_1, F_0, F_0)	π_1	$\lambda_1, \lambda_2 = \lambda_3$	Type II
S_2	(F_0, F_1, F_0)	π_2	$\lambda_1 = \lambda_2, \lambda_3$	Type I
S_3	(F_0, F_0, F_1)	π_3	$\lambda_1 = \lambda_2, \lambda_3$	Type I
S_4	(F_0, F_1, F_1)	π_4	$\lambda_1, \lambda_2, \lambda_3$	Type I
	(F_1, F_0, F_1)	π_4	$\lambda_1, \lambda_2, \lambda_3$	Type I
	(F_1, F_1, F_0)	π_4	$\lambda_1, \lambda_2, \lambda_3$	Type I
	(F_1, F_1, F_1)	π_4	$\lambda_1, \lambda_2, \lambda_3$	Type I

[a] Rate independence under each state can be illustrated by the following example: λ_0, $\lambda_1 = \lambda_2$ means that λ_0 is independent of λ_1 or λ_2.

[b] Type of functional divergence under each state.

and y_0, respectively, and υ is the internal branch length. With the phylogeny given in figure 2*B*, we have

$$f(X|\lambda; x_0) = \sum_{x_5}\sum_{x_6} P_{x_0x_5}P_{x_5x_1}P_{x_5x_2}P_{x_0x_6}P_{x_6x_3}P_{x_6x_4}$$

$$f(Y|\lambda; y_0) = \sum_{y_5}\sum_{y_6} P_{y_0y_5}P_{y_5y_1}P_{y_5y_2}P_{y_0y_6}P_{y_6y_3}P_{y_6y_4}. \quad (12)$$

The conditional probability for observing *X* and *Y* under each combined state is given by

$$\begin{aligned} f(X, Y|S_0) &= \sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0} \\ &\quad \times E[P_{x_0y_0}(\upsilon|\lambda_0) f(X|\lambda; x_0) f(Y|\lambda; y_0)] \\ f(X, Y|S_1) &= \sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0} E[P_{x_0y_0}(\upsilon|\lambda_0)] \\ &\quad \times E[f(X|\lambda; x_0) f(Y|\lambda; y_0)] \\ f(X, Y|S_2) &= \sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0} E[f(X|\lambda_1; x_0)] \\ &\quad \times E[P_{x_0y_0}(\upsilon|\lambda_0) f(Y|\lambda; y_0)] \\ f(X, Y|S_3) &= \sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0} E[P_{x_0y_0}(\upsilon|\lambda_0) f(X|\lambda; x_0)] \\ &\quad \times E[f(Y|\lambda_2; y_0)] \\ f(X, Y|S_4) &= \sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0} E[P_{x_0y_0}(\upsilon|\lambda_0)] \times E[f(X|\lambda_1; x_0)] \\ &\quad \times E[f(Y|\lambda_2; y_0)]. \end{aligned} \quad (13)$$

Therefore, the joint probability of *X* and *Y* can be generally expressed as follows:

$$p(X, Y) = \sum_{j=0}^{m-1} \pi_j f(X, Y|S_j), \quad (14)$$

where $m = 5$ in this case. As above, maximization of the likelihood $L = \Pi_k p(X^{(k)}, Y^{(k)})$ can be achieved by either the Newton-Raphson or the EM algorithm.

Next, we show how equations (13) and (14) are related to calculate the coefficients of type I and type II functional divergence. Since type II functional divergence results in no altered functional constraints between two clusters, it can be interpreted as the functional divergence configuration $S_1 = (F_1, F_0, F_0)$; i.e., cluster 0 is under F_1, but clusters 1 and 2 are under F_0. Therefore, the coefficient of type II functional divergence can be defined as $\theta_{II} = P(S_1) = P(F_1, F_0, F_0) = \pi_1$. On the other hand, type I functional divergence means that at least either cluster 1 or cluster 2 should be under F_1, regardless of the status of cluster 0. According to table 2, the coefficient of type I functional divergence is given by $\theta_I = \pi_2 + \pi_3 + \pi_4$. Moreover, if the coefficient of overall functional divergence is defined as $\pi_f = 1 - P(S_0) = 1 - \pi_0$, we have

$$\theta_I + \theta_{II} = 1 - \pi_0. \quad (15)$$

Thus, π_0 can be called the coefficient of functional constraint of the gene family.

Predicting Critical Residues for Functional Divergence

It is of great interest to (statistically) predict which sites are likely to be responsible for these type I and type II functional differences. Indeed, these sites can be further tested by experimentation, using molecular, biochemical, or transgenic approaches. We will develop site-specific profiles for this purpose, which can be achieved by the posterior analysis.

Type I Functional Divergence Predicted from the Subtree Likelihood

For the simple two-cluster case, there are only two nondegenerate states: S_0 and S_1. We wish to know the probability of S_1 for a given site when the amino acid configuration (X, Y) is observed, i.e., $P^*(S_1|X, Y)$. The prior probability of S_1 is $P(S_1) = \pi_1 = \theta_{12}$. According to the Bayesian law, we have

$$P^*(S_1|X, Y) = \frac{\theta_{12} f^*(X, Y|S_1)}{p^*(X, Y)} \quad (16)$$

where $f^*(X, Y|S_1)$ and $p^*(X, Y)$ are given by equations (4) and (5), respectively.

Next, we consider the case of multiple clusters. Similar to equation (16), when the amino acid configuration (X) at a site is given, the posterior probability of each functional divergence configuration S_i can be generally expressed as follows:

$$P^*(S_i|X) = \frac{\pi_i f^*(X|S_i)}{\sum_{j=0}^{m-1} \pi_j f^*(X|S_j)}, \qquad i = 0, 1, \ldots, m-1, \quad (17)$$

where $\pi_i = P(S_i)$ is the prior probability of state S_i. When $n = 2$, equation (17) is reduced to equation (16).

For a large gene family, with little knowledge about its functional diversity, a site-specific measure for the overall level of type I functional divergence at each site is useful. Since the coefficient of overall functional divergence of a gene family is defined as $\pi_f = 1 - \pi_0$,

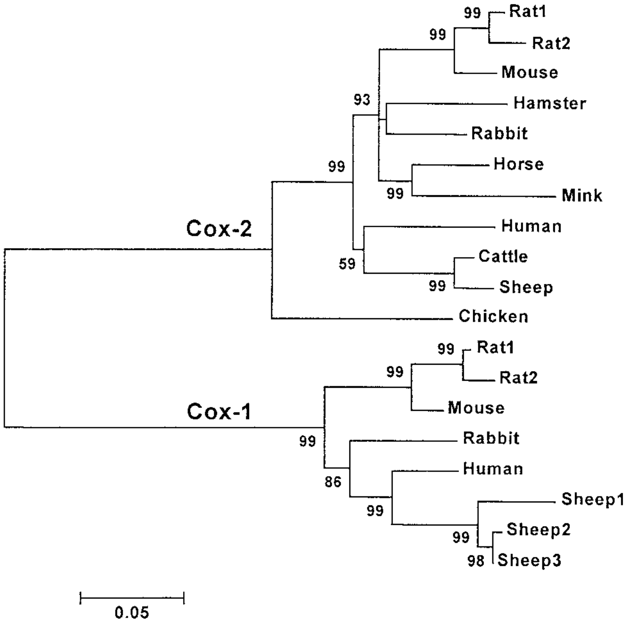

FIG. 3.—The phylogenetic tree of the COX family, which was inferred by the neighbor-joining method (Saitou and Nei 1987) using amino acid sequences with Poisson distance. Bootstrap values >50% are presented.

gwhere $\pi_0 = P(S_0)$ for $S_0 = (F_0, \ldots, F_0)$, it is natural to define such a measure as $P^*(\mathbf{F}_1|X) = 1 - P^*(S_0|X)$. According to equation (17), it is given by

$$P^*(\mathbf{F}_1|X) = 1 - \frac{\pi_0 f^*(X|S_0)}{\sum_{j=0}^{m-1} \pi_j f^*(X|S_j)}. \qquad (18)$$

Type I and Type II Functional Divergence

Based on the whole-tree likelihood for functional divergence, we can develop a site-specific profile for type I, as well as type II, functional divergence. In the case of two clusters, the posterior probability of each (nondegenerate) combined state S_i (table 2) can be computed as

$$P(S_i|X, Y) = \frac{\pi_i f(X, Y|S_i)}{\sum_{j=0}^{m-1} \pi_j f(X, Y|S_j)}, \qquad i = 0, 1, \ldots, 4, \qquad (19)$$

where $\pi_i = P(S_i)$, and $f(X, Y|S_j)$ is given by equation (13). Thus, one can easily show that site-specific profiles for type I and type II functional divergence are given by

$$\begin{aligned} P(\text{type I}|X, Y) &= P(S_2|X, Y) + P(S_3|X, Y) \\ &\quad + P(S_4|X, Y) \\ P(\text{type II}|X, Y) &= P(S_1|X, Y), \end{aligned} \qquad (20)$$

respectively.

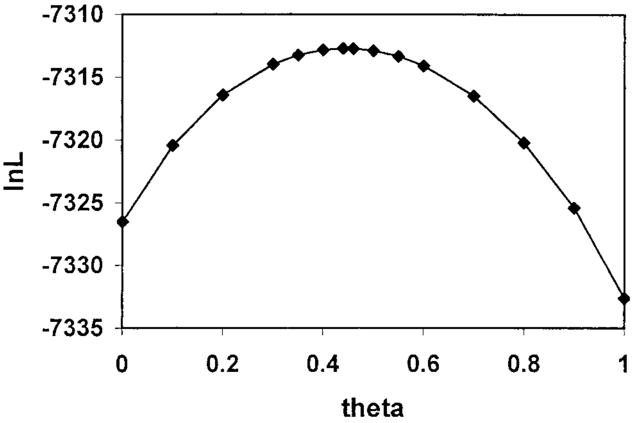

FIG. 4.—The log likelihood value of the COX gene family amino acid sequences plotted against the coefficient of functional divergence (theta).

Case Study: COX (Cyclooxygenase) Gene Family

The cyclooxygenase (COX) enzymes catalyze a key step in the conversion of arachidonate to PGH2, the immediate substrate for a series of cell prostaglandin and thromboxane synthases. Prostaglandins play critical roles in numerous biological processes, including the regulation of immune function, kidney development, reproductive biology, and gastrointestinal integrity (Williams, Mann, and DuBois 1999). There are two tissue-specific isoforms in mammals: COX-1 and COX-2. Molecular cloning of COX-2 led to a major investment by pharmaceutical companies in the development of selective inhibitors (Wallace 1999). The central tenets are that prostaglandins that contribute to inflammation are derived from COX-2, whereas prostaglandins that are involved in normal physiological processes are derived from the constitutively expressed isoform COX-1. Therefore, investigating the pattern of functional divergence between COX-1 and COX-2 amino acid sequences is helpful in drug design.

Figure 3 shows the phylogenetic tree of the COX gene family, inferred by the neighbor-joining method (Saitou and Nei 1987). It is clear that these two isoforms were generated in the early stage of vertebrates (Zou et al. 1999). Given this topology (fig. 3), the ML estimate of the coefficient of type I functional divergence between COX-1 and COX-2 is $\hat{\theta}_{12} = 0.44 \pm 0.09$, and that of the gamma shape parameter is $\hat{\alpha} = 0.43 \pm 0.05$. We conducted the LRT for the null hypothesis H_0: $\theta_{12} = 0$. The logarithm of likelihood under functional divergence is $\ln L = -7{,}312.70$, and that under H_0 ($\theta_{12} = 0$) is $\ln L = -7{,}326.51$, resulting in $\delta \ln L = -7{,}312.70 - (-7{,}326.51) = 13.81$. Thus, assuming that $2\delta \ln L$ approximately follows a χ^2 distribution, we conclude that the null hypothesis (no functional divergence) can be rejected statistically ($P < 0.01$). Figure 4 shows the shape of the log of likelihood with respect to θ_{12}.

The site-specific profile based on the posterior analysis for scoring amino acid residues that are likely to be involved in type I functional divergence between COX-1 and COX-2 is presented in figure 5. The two highest posterior probabilities ($P^*(S_1|X)$) for functional diver-

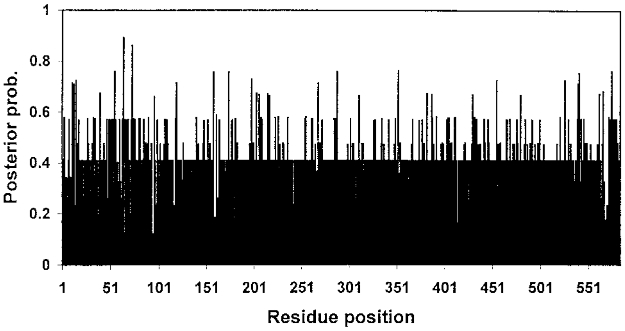

FIG. 5.—The posterior probability for predicting critical amino acid residues responsible for the functional divergence between COX-1 and COX-2 isoforms.

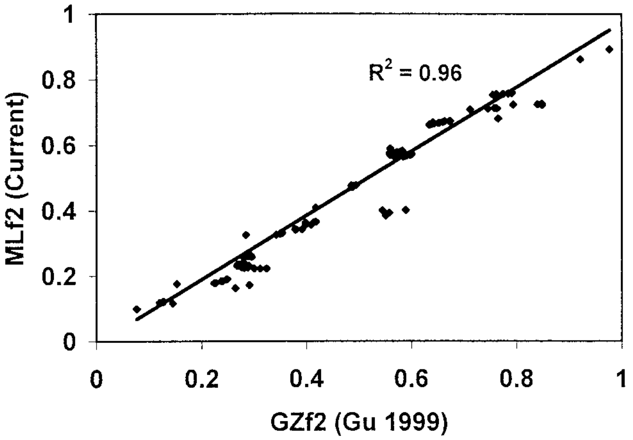

FIG. 7.—The correlation of two site-specific profiles (posterior probabilities), which are computed by the current method and Gu's (1999) method.

gence are at positions 64 and 73 (0.89 and 0.86, respectively). We found that at these sites, amino acid residues are highly conserved in COX-2 but very variable in COX-1. Figure 6 lists all amino acid residues with $P^*(S_1|X) \geq 0.70$.

For comparison, Gu's (1999) method has been applied for the same data set, which is based on a simple model but is computationally fast. Interestingly, the results from Gu's (1999) method are very close to what we obtained with the current ML approach, i.e., $\theta_{12} = 0.46 \pm 0.08$. The correlation of HMM predictions between these two methods has been studied. As shown in figure 7, these two site-specific profiles are quite similar ($R^2 = 0.96$). Indeed, amino acid residues responsible for functional divergence can be predicted by both methods (fig. 6).

Discussion

In this paper, we studied two likelihood approaches (i.e., the subtree and the whole-tree approaches) for estimating type I and type II functional divergence after gene duplication, which can be extended to the case of a large gene family with many member genes. Posterior-based site-specific profiles are implemented to predict critical amino acid residues that are responsible for functional divergence of the gene family.

The Markov Chain Model for Amino Acid Substitutions

More complicated models for amino acid substitutions can be incorporated into our framework. For example, Bruno (1996) proposed a likelihood that allows specification of the amino acid usage at a residue. Another interesting case to consider is that of the secondary structure of proteins (e.g., Thorne, Goldman, and Jones 1996). The difficulty in implementing these models in

A

```
                      1111234555
                1115671579655247
                1355439848825605
      | Sheep2  DGPITIDADGQGGENG
      | Sheep3  DGPITIDADGQGGENG
      | Sheep1  DGPSRSDADGQGGENG
Cox-1 | Human   DGPINTDADGQGGENA
      | Rabbit  DGPVSVDADGQSGENA
      | Rat1    DGPISTDIGGQRGENV
      | Rat2    DGPISTDIAGQRGENV
      | Mouse   DGPINTDVAGQRGENV

      | Hamster AAGTLVGSDQRNLHQS
      | Rabbit  AASTLVNSDLQNLNKS
      | Rat1    AASTLVGSDQRNSQRS
      | Rat2    CGSTLVGSDQRNSQRS
      | Mouse   AGSTLVGSDHQNSQKS
Cox-2 | Horse   AAGTLVGSDPRNRDKS
      | Mink    SSPTLVASDHRNSDKS
      | Cattle  AASTLVSSDFKNLEKS
      | Sheep   AVGTLVSSDIKNLEKS
      | Human   VASTLVGSDHRNMADS
      | Chicken LAGTLVSSDHQNMEES
```

B

	P(S1|X)	
Site	ML	Gu-99
64	0.893	0.977
73	0.862	0.921
352	0.763	0.791
575	0.762	0.791
288	0.759	0.791
55	0.759	0.784
158	0.758	0.774
174	0.757	0.762
198	0.728	0.848
15	0.725	0.840
526	0.725	0.794
455	0.724	0.849
11	0.714	0.762
540	0.714	0.758
119	0.714	0.763
268	0.713	0.746
13	0.710	0.711

FIG. 6.—The amino acid configurations of those residues with $P^*(S_1|X) \geq 0.70$.

practice concerns the computational complexity. In the case of the COX gene family (11 sequences, 538 residues) with two gene clusters, the current likelihood method takes about 15 min for MLE searching and posterior computation on a Pentium III PC. It can be expected that a more complex model will require more CPU time. We will investigate how much our estimation and prediction can be affected by the model selection, which can be performed with computer simulations.

Evolutionary rate and the pattern of amino acid substitutions are two fundamental components of an exploration of function/structure relationship from sequences (Lichtarge, Bourne, and Cohen 1996; Livingstone and Barton 1996; Landgraf, Fischer, and Eisenberg 1999). The current model can be extended by assuming that the pattern of amino acid substitutions differs between states F_1 and F_0. For example, a typical Dayhoff-like model is used under state F_0, while an equal-rates model is used under state F_1 that allows more radical amino acid substitutions.

The Model for Functional Divergence

Gu (1999) developed a simple method for estimating the coefficient of type I functional divergence between two gene clusters which is much faster than the current likelihood method. As shown by the example of the COX gene family, these methods give similar results. This conclusion has been verified by several other gene families (results not shown). Although extensive study is needed, it seems that estimation and prediction of functional divergence of a gene family is robust. Moreover, Gu (2000) has studied a rate-component (digamma) model for functional divergence after gene duplication. The relationship between this model and the current two-state model is discussed in appendix B.

The two-state model for functional divergence can be modified by taking the universally conserved site (type 0 in fig. 1), at which the amino acid type is invariable among all sequences of the gene family, into account. These sites are important for maintaining the common function of the gene family and/or the backbone of the three-dimensional protein structure and apparently are not involved in functional diversity among duplicate genes. Thus, we may assign a permanent F_0 state for a type 0 site in each cluster, such that the functional divergence configuration is always $S_0 = (F_0, \dots, F_0)$.

The Subtree Versus the Whole-Tree Likelihood

We have developed two likelihood frameworks. As the whole-tree likelihood (eq. 14) is built on the basis of the conventional approach (Felsenstein 1981), the subtree likelihood (eq. 4) neglects the evolutionary relationship among gene clusters. The subtree likelihood can be built for independent prior frequencies of amino acids (f^*), which is suitable for large internal branch lengths, or for identical priors (f^0), which is suitable for short internal branch lengths (see appendix A for details). Although the estimation bias caused by these approximations has been shown to be small (unpublished data), further study is needed.

Interestingly, we show in equation (A.6) that the whole-tree likelihood can be expressed as the average of f^* and f^0; the weighting function is determined by the internal branch length. Our preliminary analysis in a simple case has shown virtually no bias except for very skewed amino acid frequencies (not shown). Nevertheless, a simple solution for the potential bias correction is that the final estimate of θ is the average of θ^* and θ^0, i.e., $\hat{\theta} = (1 - e^{-\upsilon})\theta^* + e^{-\upsilon}\theta^0$, where θ^* and θ^0 are the estimates based on subtree likelihood f^* and f^0, respectively, and υ is the internal branch length. Another approach is to use ancestral sequence interference (e.g., Zhang and Nei 1997). Of course, the subtree likelihoods cannot be used to infer the evolutionary relationship among clusters or locate the ancestral node of each cluster. Although the whole-tree likelihood is apparently superior, the subtree likelihood is useful in practice for the following reasons: (1) the computation is fast for multiple gene clusters; (2) it is still applicable when the phylogenetic relationship among gene clusters is unreliable; and (3) the statistical modeling is simple.

The Assumption of Site Independence

The assumption of site independence in our modeling can be criticized because it seems to be contradicted with the great deal of evidence that covariation of residues in evolution is related to structural or functional change (e.g., Pollock, Taylor, and Goldman 1999). Here, we show that our method is actually the complement to the model of covariation, and a comprehensive analysis combining these two methods is important.

As indicated before, coevolution between sites is one important mechanism for functional divergence after gene duplication. If the evolutionary innovation occurred mainly in the early stage (e.g., cluster 0 in fig. 2), it is likely that protein evolution within each gene clusters (the late stage) may simply be under purifying selection, although functional constraints can be different between them. Since the subtree likelihood for type I functional divergence considers only gene clusters (late stages), the assumption of site independence seems reasonable.

For the whole-tree likelihood for estimating type I and type II functional divergences, modeling the mechanism of evolution in the early stage (cluster 0 in fig. 2) is achievable, e.g., a site dependence model (Pollock, Taylor, and Goldman 1999) or positive selection (Zhang, Rosenberg, and Nei 1998). The difficulty concerns the statistical identifiability and the computational feasibility. While a simple model may provide a conservative test for type II functional divergence, we will address these problems in a future study.

Effects of Number of Sequences and Phylogeny

In order to obtain reliable estimates, we suggest that each gene cluster should contain at least four sequences with sufficient sequence diversity. It is not required to use the same set of species in each gene cluster. However, our preliminary result indicates that the estimation could be very biased in an extreme case of unbalanced trees. Moreover, we found that the accuracy of posterior prediction was satisfactory as long as the phylogeny of each cluster was largely correct.

In summary, the newly developed method, as well as previous work (Gu 1999), provides a novel approach to exploring the pattern of functional divergence among member genes of a gene family. Predictions based on posterior analysis can be powerful and cost-effective in defining a group of amino acid residues that are responsible for these functional differences. Our method, combined with other approaches (e.g., positive selection analysis, coevolution of sites, ancestral inference, etc.), as well as substantial functional/structural information through high throughput technologies, is helpful in achieving a better understanding of functional diversity at the genome level.

Acknowledgments

This work was supported by NIH grant RO1 GM62118 to X.G. I am grateful to Yufeng Wang for assistance. I thank Dr. Y.-X. Fu and one reviewer for constructive comments that improved the manuscript greatly. A computer program is available on request via e-mail.

APPENDIX A
Two Types of Subtree Likelihoods

For clarity, we rewrite equation (11) here:

$$f(X, Y|\lambda) = \sum_{x_0,y_0} b_{x_0} P_{x_0y_0}(\upsilon) f(X|\lambda; x_0) f(Y|\lambda; y_0), \tag{A.1}$$

where $f(X|\lambda; x_0)$ and $f(Y|\lambda; y_0)$ are the likelihood functions for clusters 1 and 2, conditional on the roots x_0 and y_0, respectively, and υ is the internal branch length. One can show that the subtree likelihood under the assumption of independent prior of amino acids, i.e., $f^*(X, Y|\lambda) = f(X|\lambda) \times f(Y|\lambda)$, can be expressed as

$$f^*(X, Y|\lambda) = \sum_{x_0,y_0} b_{x_0} b_{y_0} f(X|\lambda; x_0) f(Y|\lambda; y_0). \tag{A.2}$$

Similarly, for the subtree likelihood under the assumption of identical priors of amino acids, it is given by

$$f^0(X, Y|\lambda) = \sum_{x_0} b_{x_0} f(X|\lambda; x_0) f(Y|\lambda; y_0). \tag{A.3}$$

Moreover, we assume the transition probability in the internal branch follows

$$\begin{aligned} P_{x_0y_0} &= b_{y_0}(1 - e^{-\upsilon}), \qquad x_0 \neq y_0 \\ P_{x_0x_0} &= b_{x_0} + (1 - b_{x_0})e^{-\upsilon}; \end{aligned} \tag{A.4}$$

i.e., the amino acid version of Felsenstein's (1981) model. Thus, we have

$$\begin{aligned} &f(X, Y|\lambda) - f^*(X, Y|\lambda) \\ &\quad = \sum_{x_0,y_0} b_{x_0}(P_{x_0y_0} - b_{y_0}) f(X|\lambda; x_0) f(Y|\lambda; y_0) \\ &\quad = -e^{-\upsilon} \sum_{x_0 \neq y_0} b_{x_0} b_{y_0} f(X|\lambda; x_0) f(Y|\lambda; y_0) \\ &\qquad + e^{-\upsilon} \sum_{x_0} (b_{x_0} - b_{x_0}^2) f(X|\lambda; x_0) f(Y|\lambda; y_0) \\ &\quad = -e^{-\upsilon} f^*(X, Y|\lambda) + e^{-\upsilon} f^0(X, Y|\lambda). \end{aligned} \tag{A.5}$$

Therefore, we have obtained an interesting result:

$$\begin{aligned} f(X, Y|\lambda) = (1 - e^{-\upsilon}) f^*(X, Y|\lambda) \\ + e^{-\upsilon} f^0(X, Y|\lambda). \end{aligned} \tag{A.6}$$

APPENDIX B
Relation to the Rate-Component Model

Gu (unpublished data) developed a rate-component model for functional divergence after gene duplication. Thus, it is interesting to investigate the relationship between the rate-component model and the two-state model. For simplicity, we consider the case of two gene clusters (fig. 1).

Let λ_1 and λ_2 be the evolutionary rates of clusters 1 and 2, respectively. The rate-component model assumes that λ_1 and λ_2 can be formulated by the following linear equations:

$$\lambda_1 = (u_0 + u_1)/\beta_1 \qquad \lambda_2 = (u_0 + u_2)/\beta_2, \tag{B.1}$$

where the rate components u_0, u_1, and u_2 are independent random variables, and β_1 and β_2 are constants. In this three-rate-component model, u_0 is the common rate component that describes the functional overlapping of two duplicate genes, whereas the independent rate component u_1 (or u_2) describes the functional constraint specific to cluster 1 (or cluster 2). Furthermore, we assume that each rate component follows a standard gamma distribution, i.e., $\pi(u_i) = u_i^{\gamma_i - 1} e^{-u_i}/\Gamma(\gamma_i)$, $i = 0, 1, 2$. Their means are $E[u_i] = \gamma_i$, and variances are $\mathrm{Var}(u_i) = \gamma_i$, respectively.

Then, one can show that λ_1 (λ_2) is also gamma-distributed, with shape parameter α_1 (or α_2) and scale parameter β_1 (or β_2), respectively. Moreover, $\alpha_1 = \gamma_0 + \gamma_1$, and $\alpha_2 = \gamma_0 + \gamma_2$. We note that the model for rate variation among sites (e.g., Gu, Fu, and Li 1995) is also a special case of the rate-component model when $u_1 = 0$ and $u_2 = 0$. Gu (2000) showed that λ_1 and λ_2 jointly follow a di-gamma distribution and obtained the analytical form of the density $\phi(\lambda_1, \lambda_2)$. In practice, however, it is more convenient to use the joint density of rate components, since they are independent, i.e.,

$$\phi(u_0, u_1, u_2) = \pi(u_0)\pi(u_1)\pi(u_2). \tag{B.2}$$

Therefore, for two-gene clusters, we have the joint probability $p(X_1, X_2)$ as follows:

$$
\begin{aligned}
p(X_1, X_2) &= \int_0^\infty \int_0^\infty \int_0^\infty f(X_1 \mid u_0, u_1) f(X_2 \mid u_0, u_2) \\
&\qquad \times \pi(u_0)\pi(u_1)\pi(u_2)\, du_0\, du_1\, du_2 \\
&= E_u[\,f(X_1 \mid u_0, u_1) f(X_2 \mid u_0, u_2)], \qquad \text{(B.3)}
\end{aligned}
$$

where the conditional probability $f(X_1 \mid u_0, u_1)$ (or $f(X_2 \mid u_0, u_2)$) can be computed according to the Markov chain property.

The relationship between the two-state model and the three-component model can be described as follows. Equation (A.1) implies that the covariance between λ_1 and λ_2 is proportional to the variance of u_0, i.e., $\mathrm{Cov}(\lambda_1, \lambda_2) = \mathrm{Var}(u_0)/\beta_1\beta_2 = \gamma_0^2/\beta_1\beta_2$. Since $\mathrm{Var}(\lambda_i) = \alpha_i/\beta_i^2$ (i = 1, 2), one can verify that the coefficient of correlation between λ_1 and λ_2 can be written as

$$r_{12} = \frac{\gamma_0}{\sqrt{\alpha_1\alpha_2}}. \qquad \text{(B.4)}$$

Since Gu (1999) has shown that under the two-state model, the coefficient of rate correlation is given by $r_{12} = 1 - \theta_{12}$, we have

$$\theta_{12} = 1 - \frac{\gamma_0}{\sqrt{\alpha_1\alpha_2}}. \qquad \text{(B.5)}$$

LITERATURE CITED

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

BRUNO, W. J. 1996. Modeling residue usage in aligned protein sequences via maximum likelihood. Mol. Biol. Evol. **13**: 1368–1374.

CASARI, G., C. SANDER, and A. VALENCIA. 1995. A method to predict functional residues in proteins. Struct. Biol. **2**:171–178.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. Pp. 342–352 *in* M. O. DAYHOFF, ed. Atlas of Protein Sequence Structure. vol 5, Suppl. 3. National Biomedical Research Foundation, Washington, D.C.

DEMPSTER, A. P., N. M. LAIRD, and D. B. RUBIN. 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39**:1–38.

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

GOLDING, G. B., and A. M. DEAN. 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., 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., 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.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**:275–282.

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

KISHINO, H., T. MIYATA, and HASEGAWA. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

LANDGRAF, R., D. FISCHER, and D. EISENBERG. 1999. Analysis of heregulin symmetry by weighted evolutionary tracing. Protein Eng. **12**:943–951.

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.

LICHTARGE, O., H. R. BOURNE, and F. E. COHEN. 1996. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. **257**:342–358.

LIVINGSTONE, C. D., and G. BARTON. 1996. Identification of functional residues and secondary structure from protein sequence alignment. Methods Enzymol. **266**:497–512.

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.

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

POLLOCK, D., W. R. TAYLOR, and N. GOLDMAN. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. **287**:187–198.

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.

THORNE, J. L., N. GOLDMAN, D. T. JONES. 1996. Combining protein evolution and secondary structure. Mol. Biol. Evol. **13**:666–673.

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

WALLACE, J. L. 1999. Distribution and expression of cyclooxygenase (COX) isoenzymes, their physiological roles, and the categorization of nonsteroidal antiinflammatory drugs. Am. J. Med. **107**:11S–16S.

WILLIAMS, C. S., M. MANN, and R. N. DUBOIS. 1999. The role of cyclooxygenases in inflammation, cancer, and development. Oncogene **18**:7908–7916.

WOLFE, K. H., and D. C. SHIELDS. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**:708–713.

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

———. 1994. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites: approximate methods. J. Mol. Evol. **39**:306–314.

ZHANG, J., and M. NEI. 1997. Accuracy of ancestral amino acid sequences inferred by the parsimony, likelihood and distance method. J. Mol. Evol. **44**:S139–S146.

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.

ZOU, J., N. F. NEUMANN, J. W. HOLLAND, M. BELOSEVIC, C. CUNNINGHAM, C. J. SECOMBES, and A. F. ROWLEY. 1999. Fish macrophages express a cyclo-oxygenase-2 homologue after activation. Biochem. J. **340**:153–159.

YUN-XIN FU, reviewing editor

Accepted November 29, 2000