A Simple Statistical Method for Estimating Type-II (Cluster-Specific) Functional Divergence of Protein Sequences

Xun Gu

doi:10.1093/molbev/msl056

Predicting functional amino acid residues in silico is important for comparative genomics. In this paper, we focus on the issue of how to statistically identify cluster-specific amino acid residues that are related to the functional divergence after gene duplication. We approach this problem using a framework based on site-specific shift of amino acid property (type-II functional divergence), as opposed to site-specific shift of evolutionary rate (type-I functional divergence). An efficient statistical procedure is implemented to facilitate the development of phylogenomic database for cluster-specific residues of large-scale protein families. Our method has the following features: 1) statistical testing of the type-II functional divergence and 2) the site-specific Bayesian profile to measure how amino acid residues contribute to type-II (cluster-specific) functional divergence. Consequently, one may obtain the posterior probability for “functional” cluster-specific residues. Case studies are presented and indicate that radical cluster-specific residues are responsible for most of inferred type-II functional divergence, whereas conserved cluster-specific residues appear less than even those imperfect radical cluster-specific residues to this type of functional divergence.

Introduction

Under the framework of phylogenomic annotation of gene function (Eisen and Fraser 2003), the importance of gene function can be measured quantitatively in terms of the functional constraints of the protein sequence (Kimura 1983). For instance, an amino acid residue is said to be functionally important if it is evolutionarily conserved. Therefore, change of the evolutionary conservation at a particular residue may indicate the involvement of functional divergence (Lichtarge et al. 1996; Gu 1999). Following this idea, many research groups have developed statistical methods for testing and predicting the functional divergence of a gene family, which indeed showed the association between sequence and functional or structural divergence (e.g., Lichtarge et al. 1996; Gu 1999, 2001, 2003; Knudsen and Miyamoto 2001; Landgraf et al. 2001; Wang and Gu 2001; Gaucher et al. 2002; Jordan et al. 2001; Lopez et al. 1999; Gribaldo et al. 2003; Madabushi et al. 2004; Gao et al. 2005; Rastogi and Liberles 2005; H Zhou, J Gu, SJ Lamont, X Gu, personal communication).

Furthermore, Gu (2001) made a distinction between 2 types of functional divergence. Type-I functional divergence results in site-specific rate shift (Gu 1999; Knudsen and Miyamoto 2001; Landgraf et al. 2001; Gaucher et al. 2002; Lopez et al. 2002). A typical case is an amino acid residue that is highly conserved in a subset of homologous genes but highly variable in a different subset of those homologous genes. Alternatively, type-II functional divergence results in the shift of cluster-specific amino acid property (Lichtarge et al. 1996; Gu 2001). Such divergence is exemplified by a radical shift of amino acid property, for example, positive versus negative charge differences at a homologous site that is otherwise evolutionally conserved between subtrees within a phylogeny. Note that these 2 types of functional divergence may have other names. For instance, the basic evolutionary trace approach (Lichtarge et al. 1996; Madabushi et al. 2004) has mainly focused on cluster-specific residues related to type-II functional divergence. Gribaldo et al. (2003) also looked at type-II functional divergence called as “constant-but-different.” Meanwhile, the weighted evolutionary trace approach proposed by Landgraf et al. (2001) was similar to type-I functional divergence (Gu 1999).

Many studies have been published about the statistical significance of observed patterns, which is important as the research community tends to use them to infer functional divergence of proteins (e.g., Gu 1999, 2001; Knudsen and Miyamoto 2001; Landgraf et al. 2001; Gaucher et al. 2002). Although several methods for type-I functional divergence are available, as well as the software (e.g., Gu and Vander Velden 2002), the implementation of statistical testing for type II has not been well resolved.

In this paper, we develop a statistical method for type-II functional divergence. In a typical case of 2 gene clusters generated by a gene duplication event, type-II functional divergence results in site-specific shift of amino acid physicochemical property after the gene duplication. In this regard, “type II is also called cluster-specific functional divergence” (Lichtarge et al. 1996). Cluster-specific amino acid residues have been widely used in functional assays of protein families. For instance, Sun et al. (2002) identified essential amino acid changes in paired domain evolution of the PAX gene family. Given the growing diversity and number of protein sequences available, it is now practical and necessary to develop a phylogenetic database for cluster-specific amino acid residues of protein families. To this end, we have to address 2 related statistical issues. First, are the type-II (cluster-specific) changes statistically significant? And secondly, for observed cluster-specific amino acid residues, how can we statistically measure whether they are related to type-II functional divergence. We address these issues by developing the statistical method that is suitable for large-scale data analysis.

Theory

Type-II Functional Divergence (cluster-specific) in the Early Stage

In principle, the evolution of protein sequences of duplicate genes can be divided into 2 stages, the early (E) stage after gene duplication and the late (L) stage (fig. 1). We
assume that functional divergence between duplicate genes has occurred in the E stage, whereas in the late (L) stage, the purifying selection plays a major role to maintain related but distinct functions of 2 duplicate genes (Ohno 1970; Kimura 1983; Force et al. 1999). Accordingly, we modify the “2-state model” (Gu 1999, 2001) specific to type-II (cluster-specific) functional divergence:

(i) In the E stage, an amino acid residue can be in either of 2 states: $F_0$ (type II unrelated) and $F_1$ (type II related). The probability of a residue being under $F_1$ is $P(F_1) = \theta_H$ and that being under $F_0$ is $P(F_0) = 1 - \theta_H$. To distinguish it from the type-I functional divergence (Gu 1999), we call $\theta_H$ “the coefficient of type-II functional divergence.”

(ii) In the L stage, an amino acid residue is always under the state of $F_0$, indicating no further type-II functional divergence. Amino acid substitutions in this stage are mainly under purifying selection.

Substitution Models under $F_0$ and $F_1$

The pattern of amino acid substitutions during evolution, or the substitution model, relies on the states of functional divergence ($F_0/F_1$). The $F_0$-substitution model largely reflects the conserved evolution of protein sequences, which can be empirically determined by the Dayhoff model (Dayhoff et al. 1978) or the Jones-Taylor-Thornton model (Jones et al. 1992). In contrast, under $F_1$, radical amino acid substitutions may occur more frequently, apparently due to the functional divergence between duplicate genes (Lichtarge et al. 1996). To avoid overparameterization, we propose a simple $F_1$-substitution model that can...
distinguish between the “radical” and “conserved” amino acid substitutions. First, we tentatively classify 20 amino acids into 4 groups: charge positive (K, R, and H), charge negative (D and E), hydrophilic (S, T, N, Q, C, G, and P), and hydrophobic (A, I, L, M, F, W, V, and Y). An amino acid substitution is called radical (denoted by $R$) if it changes from one group to another; otherwise it is called conserved, that is, within the group, denoted by $C$. It should be mentioned that the conservations of residues may be due to selection pressure for greater functions or for folding stabilities (Tseng and Liang 2006). The status of no substitution is denoted by $N$.

Secondly, we assume that, under state $F_0$, the transition probability for a radical, conserved, or no substitution is given by

\[ P(R|F_0) = \pi_R(1 - e^{-\lambda t}), \]
\[ P(C|F_0) = \pi_C(1 - e^{-\lambda t}), \]

or

\[ P(N|F_0) = e^{-\lambda t}, \]

(1)

respectively, where $t$ is the evolutionary time, $\lambda$ is the substitution rate, and $\pi_R$ (or $\pi_C$) is the proportion of radical (or conserved) substitutions in the total substitutions; $\pi_R + \pi_C = 1$. Apparently, equation (1) is an extended Poisson model of protein sequence evolution. Based on the Dayhoff PAM matrix, we empirically determined that $\pi_R = 0.312$ and $\pi_C = 0.688$. Indeed, without any functional divergence, conserved amino acid substitutions are more likely to occur, as expected by the theory of neutral evolution (Kimura 1983).

Next we consider the transition probabilities under $F_1$ in the early stage, denoted by $P(Y|F_1)$ for $Y = N, R, C$. It should be noted that, according to our model (see above), an amino acid residue that has no change in the early stage is essentially unrelated to the type-II functional divergence. This argument implies $P(N|F_1) = 0$. Similar to Eq.(1), one may choose $P(R|F_1)$ and $P(C|F_1)$ in the same forms as $a_R(1 - e^{-\lambda t})$ and $a_C(1 - e^{-\lambda t})$, respectively. Together, these arguments directly lead to

\[ P(R|F_1) = a_R, \]
\[ P(C|F_1) = a_C, \]

and

\[ P(N|F_1) = 0. \]

(2)

where $a_R = a_R(1 + a_C)$ and $a_C = 1 - a_R$. That is, $a_R$ (or $a_C$) is the ($F_1$) proportion of radical (or conserved) substitutions in total substitutions. Moreover, the $F_1$-radical amino acid substitution ($a_R$) can be much higher than that under $F_0$ ($\pi_R$), as will be shown later.

Evolutionary Link between Early and Late Stages

The evolutionary link between early and late stages depends on the status of type-II (cluster-specific) functional divergence. Let $\lambda_E$ and $\lambda_L$ be the evolutionary rates in the E and L stages, respectively. The statistical framework we developed is under the following assumptions:

(i) A random variable $u$, called the rate component, varies among sites according to a standard gamma distribution

\[ \phi(u) = \frac{\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-u}. \]

(3)

The shape parameter $\alpha$ describes the strength of rate variation among sites, that is, a small value of $\alpha$ means a strong rate heterogeneity among sites, and $\alpha = \infty$ means no rate variation among sites (Gu et al. 1995).

(ii) Under $F_0$, the evolutionary rates in the early ($\lambda_E$) and late ($\lambda_L$) stages share the same rate component $u$. That is, $\lambda_E = c_1 u$ and $\lambda_L = c_2 u$, where $c_1$ and $c_2$ are constant.

(iii) $F_1$-amino acid substitutions in the early stage are independent of the rate component $u$, as indicated by equation (2). In other words, $F_1$-amino acid substitutions have escaped from the ancestral functional constraint on the protein sequence.

Two Clusters by Gene Duplication

Consider the typical case of 2 clusters generated by a gene duplication event, each of which consists of several orthologous genes (fig. 1). Let $X$ be the amino acid pattern of the late stage, the column (site) of the multiple alignments of the sequences. Let $Y = (a, b)$ be the amino acid pattern of the early stage, the ancestral sequences of 2 internal nodes $a$ and $b$. From the assumption (ii), the joint probability of $X$ and $Y$ under $F_0$ is given by $P(X, Y|F_0) = \int_0^\infty P(X|Y)P(Y|F_0)\phi(u)du$, where $P(Y|F_0)$ is determined by equation (1) for $Y = N, C, a$, or $b$, respectively, $P(X|Y)$ is the likelihood of the subtrees of the 2 clusters $A$ and $B$, conditional on the ancestral states $a$ and $b$, which can be constructed according to the Markov-chain property under a known phylogeny (Felsenstein 1981; Gu 2001), and the gamma distribution density for rate variation among sites $\phi(u)$ is given by equation (3). Similarly, from (iii), under $F_1$ we have $P(X, Y|F_1) = P(Y|F_1) \times \int_0^\infty P(X|Y)\phi(u)du$, where $P(Y|F_1)$ is given by equation (2). Remembering that the probability of a site being under $F_1$ is given by $P(F_1) = \theta_0$, the coefficient of type-II functional divergence, we have the joint probability for $X$ and $Y$ as follows:

\[ P(X, Y) = (1 - \theta_0)P(X, Y|F_0) + \theta_0 P(X, Y|F_1). \]

(4)

Direct application of equation (4) for estimating $\theta_0$ may face some difficulties because the amino acid pattern of early stage ($Y$) is unobservable. A straightforward solution is to invoke the ancestral sequence inference, for example, Yang et al. (1995). Treating the ancestral sequences as inferred observations, the standard procedure for the likelihood analysis of protein sequence can be applied. In spite of nice statistical properties, this approach requires a detailed description of the model and sensitive to the statistical uncertainty in ancestral sequence inference. To solve this problem, we thus propose a simple but robust method that is computationally efficient, allowing genome-wide proteomic analysis.
A Simple Robust Method

Poisson Model in the Late Stage

Testing type-II functional divergence between 2 gene clusters (the early stage) utilizes the within-cluster amino acid patterns to examine the conservation in the late stage. Therefore, a Poisson-based model that counts the number of substitutions may be sufficient for this purpose, where smaller values of \( k \) of substitutions in a gene cluster indicate high conservation. Formally, at a given amino acid residue, the number of substitutions in each cluster (A or B) follows a Poisson process, for example, for cluster A, we have

\[
p_A(k) = \frac{(\lambda_A T_A)^k e^{-\lambda_A T_A}}{k!},
\]

(5)

with the same applying to \( p_B(k) \), where \( T_A \) (or \( T_B \)) is the total evolutionary time of cluster A (or B), and \( \lambda_A \) (or \( \lambda_B \)) is the evolutionary rate of cluster A (or B), respectively. Hence, the early–late joint distribution can be specified as

\[
f_{ij} = P(X = (i, j), Y), \text{ where } i \text{ or } j \text{ is the number of substitutions in cluster A or B.}
\]

Under this model, \( P(XY) = p_A p_B \), which is independent of the early stage. Similar to the derivation of equation (4), we have

\[
P(X = (i, j), Y) = \int_0^{\infty} P(Y|F_0)p_A(i)p_B(j)\,\phi(u)\,du = \int_0^{\infty} P(Y|F_1)p_A(i)p_B(j)\,\phi(u)\,du.
\]

Together, one can show that the early–late distribution under the Poisson-based model is given by

\[
f_{ij} = (1 - \theta_B) \int_0^{\infty} P(Y|F_0)p_A(i)p_B(j)\,\phi(u)\,du + \theta_B a_Y \int_0^{\infty} p_A(i)p_B(j)\,\phi(u)\,du,
\]

(6)

where \( P(Y|F_0) \) is from equation (1) and \( P(Y|F_1) \) from equation (2); here \( a_N = P(N|F_1) = 0 \).

Analytical Form of the Early–Late Distribution

First we consider the late-stage distribution, \( f_{ji} \), the probability for \( i \) and \( j \) substitutions in clusters A and B, respectively. From equation (6), one can show that

\[
f_{ij} = f_{ij,E} + f_{ij,C} = \int_0^{\infty} p_A(i)p_B(j)\,\phi(u)\,du = Q_{ij} \text{ is a specific version of bivariate negative binomial distribution},
\]

\[
Q_{ij} = \frac{\Gamma(i + j + \alpha)}{i! j! \Gamma(\alpha)} Z^i Z_\alpha^j Z_B^j.
\]

(7)

where \( Z = \alpha((D_A + D_B + \alpha), Z_A = D_A(D_A + D_B + \alpha), Z_B = D_B(D_A + D_B + \alpha); D_A = \lambda_A T_A \) and \( D_B = \lambda_B T_B \) are the total branch lengths of clusters A and B, respectively, and \( \alpha \) is the gamma shape parameter.

Next we consider the early-stage distribution \( f_i \), the frequencies of \( 3 \) early-stage amino acid patterns for \( Y = \{N, R, \text{ or } C\}. \) Because \( f_i = \sum_{j} f_{ij} \), from equation (6) one can show

\[
f_i = (1 - \theta_B)(1 - e^{-d})\,\tau_i + \theta_B a_Y, \quad Y = R, \text{ or } C
\]

(8)

and \( f_C = 1 - f_N - f_R \). Moreover, let \( p = p_R + p_C \) be the proportion of amino acid differences (either radical or conserved) in the early stage, which is given by

\[
p = (1 - \theta_B)(1 - e^{-d}) + \theta_B,
\]

(9)

where \( d = \tilde{d}T \) is the branch length of the early stage.

We define \( W = y(D_A + D_B + d + \alpha), W_A = D_A(D_A + D_B + d + \alpha), \) and \( W_B = D_B(D_A + D_B + d + \alpha) \). Finally, we have shown that the joint distribution of early–late stages, \( f_{ij} \), \( y \), can be expressed as follows:

\[
f_{ij,N} = (1 - \theta_B)M_{ij},
\]

\[
f_{ij,R} = (1 - \theta_B)(Q_{ij} - M_{ij})\,\tau_R + \theta_B a_R Q_{ij},
\]

and

\[
f_{ij,C} = (1 - \theta_B)(Q_{ij} - M_{ij})\,\tau_C + \theta_B a_C Q_{ij},
\]

(10)

where \( M_{ij} = \int_0^{\infty} e^{-s}\,p_A(i)p_B(j)\,\phi(s)\,ds \) is given by

\[
M_{ij} = \frac{\Gamma(i + j + \alpha)}{i! j! \Gamma(\alpha)} W^i W_A^j W_B^j.
\]

(11)

Estimation

Based on the likelihood principle, we have implemented the following algorithms to estimate unknown parameters for testing type-II functional divergence. Here we always assume that the phylogenetic tree of the gene family is known or can be reliably inferred.

Late-Stage Likelihood

The distribution of late-stage \( Q_{ij} \) is the probability of a site being \( i \) and \( j \) substitutions in the 2 clusters. As shown by equation (7), \( Q_{ij} \) depends on 3 (late-stage) parameters \( D_A, D_B, \) and \( \alpha \). We thus modified the likelihood method of Gu and Zhang (1997) to estimate them simultaneously, denoted by \( D_A, D_B, \) and \( \beta \), respectively. Note that the algorithm of Gu and Zhang (1997) corrected the parsimony bias in counting the number of substitutions.

Likelihood for Estimating Early-Stage Parameters

Let \( n_{ij,y} \) be the number of site with the pattern \( X = (i, j) \) and \( Y = \{N, R, \text{ or } C\}. \) After treating \( 3 \) late-stage parameters as known, we develop a simple likelihood to estimate early-stage parameters \( \theta_B, a_R, a_C, \) and \( d \). From equation (10), we have

\[
f_{ij,S} = f_{ij,E} + f_{ij,C} = Q_{ij} - (1 - \theta_B)M_{ij} = Q_{ij} - f_{ij,N}.
\]

Let \( n_{ij,S} = n_{ij,R} + n_{ij,C} \). Thus, the log-likelihood function,

\[
\ell = \sum_{ij} n_{ij,N} \ln((1 - \theta_B)) + \ln(M_{ij})
\]

(12)

\[
+ \sum_{ij} n_{ij,S} \ln(Q_{ij} - f_{ij,N}),
\]

includes 2 unknown parameters \( \theta_B \) and \( d \). Let \( N_0 = \sum_{ij} n_{ij,N} \) is the total number of sites that have no change in the early stage. Under the \( p \) constraint of equation (9), the maximum likelihood estimate of \( \theta_B \) is given by \( \theta = 1/(1 - y) \), where \( y \) is the solution of

\[
\sum_{ij} n_{ij,S} M_{ij} = N_0, \quad Q_{ij} - M_{ij} = N_0,
\]

(13)

with \( d = -\ln(1 - p) + \ln(1 - \theta_B) \). (Note that \( M_{ij} \) depends on the parameter \( d \), whereas \( Q_{ij} \) only depends on late-stage
Let \( L \) be the sequence length, with the initial values of parameters that are treated as known. The iteration can start with the initial values of \( d^0 = -\ln(1 - p) \) until convergence. Let \( L \) be the sequence length, \( f_{0j} = n_{0j}/L \) and \( f_0 = N_0/L \). The sampling variance of \( \theta_0 \) can be calculated as follows:

\[
\text{Var}(\hat{\theta}) = \frac{1}{L(f_0 + a)}
\]

where \( a = \sum_i \hat{f}_{ij} M_{ij}^2 / (Q_i - M_{ij}^2) \hat{\theta}_n^2 \). When the estimates \( \hat{\theta}_n \) and \( d \) are obtained, \( \Delta R \) can be estimated from equation (8).

The proportion of amino acid differences between the internal nodes \( a \) and \( b \) represented by \( p \) can be computed as follows. First, we use the Bayesian algorithm (Zhang and Nei 1997) to infer the ancestral sequences of \( Y \), which is a simplified version of Yang et al. (1995) in which the branch lengths of the phylogenetic tree are estimated using a least square method rather than the ML method. Then, we estimate \( p \) when each site in the inferred ancestral sequence receives the assignment of amino acid with the highest posterior probability. Simulations conducted by Zhang and Nei (1997) showed that this approach for estimating \( p \) is almost unbiased.

### The U-Likelihood

This method utilizes amino acid sites that are universally conserved in both clusters, that is, \( i = j = 0 \). Let \( n_{00} \) be the number of sites with \( U = N \) (the \( U \) type), \( R \), or \( C \). Let \( n_{00} = n_{00N} + n_{00R} + n_{00C} \) and \( f_{00} = f_{00N} + f_{00R} + f_{00C} \). Then, the log of \( U \)-likelihood can be written as

\[
\ell_u = \sum_{Y = N, R, C} n_{00Y} \ln f_{00Y} + (N - n_{00}) \ln (1 - f_{00})
\]

Let \( f_{00N} = n_{00N} / N \). Similar to above, we have shown that the ML estimates of \( \theta_0 \) and \( d \) are given by

\[
\hat{\theta}_0 = 1 - f_{00N} \left[ 1 + \frac{\hat{D}_n + \hat{H}_n + d}{2} \right]^{-1}
\]

\[
d = -\ln(1 - p) + \ln(1 - \hat{\theta}_0)
\]

The sampling variance of the estimates \( \hat{\theta}_0 \) is \( \text{Var}(\hat{\theta}_0) = f_{00N}(1 - f_{00N}) b^2 / N \), where \( b = [1 + (D_n + D_b + d)/2]^2 \). Because the \( U \) method largely relies on the universally conserved sites, it seems robust against the inaccuracy of ancestral sequence inference and sequence alignment.

### Predicting Critical Amino Acid Residues: Empirical Bayesian Approach

The identification of which sites are responsible for these type-II (cluster-specific) functional differences is of great interest, if the coefficient of functional divergence \( \theta_0 \) between early and late stages is significantly larger than 0. Here we develop a method of predicting such sites, which indeed can be further tested by experimentation, using molecular, biochemical, or transgenic approaches.

We wish to know the probability of state \( F_1 \) in the early stage at a site, that is, \( P(F_1|X, Y) \). According to the Bayesian law, we have

\[
P(F_1|X, Y) = \frac{P(F_1)P(X|Y|F_1)}{P(X|Y)},
\]

where the prior probability of \( F_1 \) in the early stage is given by \( P(F_1) = \theta_0 \). Under the Poisson-based model, \( P(X = (i, j), YF_1) \) and \( P(X = (i, j), YF_0) \), and \( P(X = (i, j), Y) \) are given by equations (5) and (7), respectively. Noting that \( a_Y = 0 \) if \( Y = N \), one can show

\[
P(F_1|X, Y) = 0 \quad \text{if} \quad Y = N,
\]

\[
P(F_1|X, Y) = a_Y \theta_0 Q_{ij}/f_{ijY} \quad \text{if} \quad Y = C,
\]

and

\[
P(F_1|X, Y) = a_R \theta_0 Q_{ij}/f_{ijY} \quad \text{if} \quad Y = R.
\]

One may find that it is simple to use the posterior probability ratio of \( F_1 \) to \( F_0 \), that is, \( R(F_1|F_0) = P(F_1|X, Y) / P(F_0|X, Y) \). After some algebras, we obtain

\[
R(F_1|F_0) = \frac{\theta_0}{1 - \theta_0} \frac{a_Y}{\pi_{NC}} \frac{1}{1 - (1 - h)^{1 + y + z}} \quad \text{if} \quad Y = N,
\]

\[
R(F_1|F_0) = \frac{\theta_0}{1 - \theta_0} \frac{a_R}{\pi_{NC}} \frac{1}{1 - (1 - h)^{1 + y + z}} \quad \text{if} \quad Y = R,
\]

where \( h = d/(D_A + D_B + d + 2) \).

### Statistical Evaluation of Cluster-Specific Sites

An important result from equation (19) is that the posterior ratio \( R(F_1|F_0) \) reaches its maximum if there is no amino acid substitution in each gene cluster but the amino acid is different between them, that is, \( i = j = 0 \) and \( Y = F \). As usually observed, and assuming that the proportion of radical changes under \( F_1 \) is higher than that under \( F_0 \) such that \( a_R/a_C > \pi_R/\pi_C \), we have

\[
R(F_1|F_0)_{\max} = \frac{\theta_0}{1 - \theta_0} \frac{a_R}{\pi_{NC}} \frac{1}{1 - (1 - h)}.
\]

Hence, a typical cluster-specific site indeed will receive a highest score for the type-II functional divergence, consistent with the intuitive biological interpretation. However, it should also be indicated that a high score could be statistically meaningless if \( \theta_0 \) is not significantly larger than 0. Finally, we note that \( R(F_1|F_0)_{\max} \rightarrow \infty \) if \( h \rightarrow 0 \). This means that greater accuracy is achieved as more sequences are analyzed (i.e., increasing \( D_A \) or \( D_B ) \). In practice, one may use this property to determine how many sequences are sufficient to achieve the statistical resolution of site prediction.

### Software and Examples

The newly developed method for type-II functional divergence has been implemented in the software package DIVERGE2, which is available from our Web site http://xgu.gdc.b.iastate.edu. We have distribution packages for both Microsoft Windows and Linux operating systems.
1942 Gu

isoforms (Gq and GS) of G-protein alpha subunits and found significant (efficients of type-I and type-II functional divergence, respectively. SE: standard error.

des a tree (see fig. 1, panel
dvidence for type-II functional divergence between 2 major
ritional divergence of caspase gene family, we found no ev-

In contrast to Wang and Gu’s (2001) finding for type-I func-
a similar pattern of type-II functional divergence (table 1).  

Data Sets
We present case studies analyzing 3 gene families. 1) The cyclooxygenase (COX) enzymes catalyze a key step in the conversion of arachidonic to PGH2, the immediate substrate for a series of cell prostaglandin and thromboxane synthases. There are 2 tissue-specific isofoms in mammals: COX1 and COX2. Molecular cloning of COX2 led to a major investment by pharmaceutical companies in the development of selective inhibitors. The sequence alignment and the phylogenetic tree are the same as in Gu (2001); also see figure 1C. 2) The caspase gene family is important for apoptosis (programmed cell death) and cytokine maturation, which has been studied extensively for type-I functional divergence (Wang and Gu 2001). And 3) we have also analyzed the duplicated isofoms (Gq and GS) of G-protein alpha subunits.

Testing the Significance of Type-II Functional Divergence
As expected, all these gene families show significant type-I functional divergence. Based on the phylogenetic tree in figure 1C, we estimated that the coefficient of type-II functional divergence between COX1 and COX2 duplicate genes \( \theta_{II} = 0.159 \pm 0.036 \), which is statistically significant (\( P < 0.001 \)). We also analyzed the duplicated isofoms (Gq and GS) of G-protein alpha subunits and found a similar pattern of type-II functional divergence (table 1). In contrast to Wang and Gu’s (2001) finding for type-I functional divergence of caspase gene family, we found no evidence for type-II functional divergence between 2 major

including manual and example files. We have conducted several case studies to demonstrate its potential applications in understanding functional divergence of protein sequences.

Site-Specific Profiles of Type-II (cluster-specific) Functional Divergence
For illustration, figure 2 shows the site-specific ratio profile of type-II functional divergence between COX1 and COX2. For 583 aligned sites, 492 (84%) sites have received the ratio score \( < 1 \), indicating that most sites are predicted to be unrelated to the type-II functional divergence. Moreover, we identified 28 radical cluster-specific sites that receive the highest posterior ratio score, that is, \( R(F_1 F_0)_{\text{max}} = 7.17 \). In other words, if we select these radical cluster-specific sites as candidates for type-II functional divergence, the posterior probability for them is \( P_{II} = 7.17/(1 + 7.17) = 87.8\% \), indicating that the prediction error (false-positive rate) is 12.2%. Actually, it is impressive that among 111 radical amino acid substitutions in the early stage after the gene duplication, about 29/111 \( \approx 26\% \) are potentially related to type-II functional divergence between COX1 and COX2.

Effect of Radical Substitutions in the Early Stage
For the COX gene family, we found that the radical substitutions for type-II functional divergence in the early stage is about 2.7-fold increasing (\( a_{II}/R = 2.7 \) in table 1). Consequently, an amino acid residue with a radical change between COX1 and COX2 may have a higher score than a conserved change for being type-II functional divergence related. As shown by table 2, the sites most likely exhibiting type-II behavior are the radical cluster-specific sites, whereas the conserved cluster-specific sites are less likely, as indicated by a low posterior probability (\( \approx 0.35 \)). This case study clearly shows the important role of statistical analysis, otherwise one cannot objectively justify whether one-less radical cluster-specific site (i.e., there is one amino acid substitution in the late stage) is more likely to be functional divergence related than conserved cluster-specific site.

Table 1

<table>
<thead>
<tr>
<th>Type II</th>
<th>COX</th>
<th>G-Protein Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>370</td>
<td>151</td>
</tr>
<tr>
<td>C</td>
<td>102</td>
<td>72</td>
</tr>
<tr>
<td>R</td>
<td>111</td>
<td>111</td>
</tr>
<tr>
<td>( \rho )</td>
<td>0.365</td>
<td>0.548</td>
</tr>
<tr>
<td>( D_a )</td>
<td>0.376</td>
<td>0.820</td>
</tr>
<tr>
<td>( D_b )</td>
<td>0.590</td>
<td>0.944</td>
</tr>
<tr>
<td>( d )</td>
<td>0.282</td>
<td>0.402</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>0.401</td>
<td>0.440</td>
</tr>
<tr>
<td>( f_a )</td>
<td>0.521</td>
<td>0.607</td>
</tr>
<tr>
<td>( a_{II}/R )</td>
<td>2.744</td>
<td>2.811</td>
</tr>
<tr>
<td>( \theta_{II} + \text{SE} )</td>
<td>0.159 ± 0.036</td>
<td>0.325 ± 0.055</td>
</tr>
</tbody>
</table>

Note.—\( N, C, \) and \( R \) are the numbers of sites across internal nodes \( (a, b) \) of the tree (see fig. 1, panel A) that display no difference, conserved difference, and radical differences, respectively, and \( \rho \) is the proportion of (overall) differences between nodes \( a \) and \( b \). \( D_a \) and \( D_b \) are the average numbers of substitutions per sites in clusters \( A \) and \( B \), respectively, and \( d \) is the distance between nodes \( a \) and \( b \). The parameter \( \alpha \) is the gamma shape parameter. \( f_a \) is the observed proportion of radical changes in all substitutions between nodes \( a \) and \( b \). \( a_{II}/R \) is the ratio of radical changes under (type-II) functional divergence versus nonfunctional divergence. Finally, \( \theta_{II} \) and \( \theta_{II} \) are the coefficients of type-I and type-II functional divergence, respectively. SE: standard error.
Table 2

<table>
<thead>
<tr>
<th>Functional Ranking of Several Cluster-Specific Patterns in the COX Gene Family</th>
<th>Between Clusters (early stage)</th>
<th>Within Clusters (late stage)</th>
<th>Number of Sites</th>
<th>Ratio Score</th>
<th>Posterior Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Radical change (radical cluster specific)</td>
<td>No a.a. change</td>
<td>28</td>
<td>7.17</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>(2) Radical change</td>
<td>One a.a. change</td>
<td>30</td>
<td>2.11–2.22</td>
<td>0.68–0.69</td>
<td></td>
</tr>
<tr>
<td>(3) Radical change</td>
<td>Two a.a. changes</td>
<td>20</td>
<td>1.25–1.41</td>
<td>0.56–0.59</td>
<td></td>
</tr>
<tr>
<td>(4) Conserved change (conserved cluster specific)</td>
<td>No a.a. change</td>
<td>31</td>
<td>0.55</td>
<td>0.35</td>
<td></td>
</tr>
</tbody>
</table>


Discussion

The importance of cluster-specific amino acid residues for understanding functional divergence of a protein family has been well recognized (Lichtarge et al. 1996; Sun et al. 2002). Because the computational prediction of these sites from a multiple alignment is straightforward, it is technically simple to develop a phylogenomic database for cluster-specific residues of large-scale protein families. The difficulty, however, is the statistical issue. Without developing a reliable statistical procedure to test the significance of observed cluster-specific residues, such phylogenomic effort could result in rapid accumulation of false-positive cases. Based on the statistical framework of type-II functional divergence, we have provided a practical solution for this problem.

We first test type-II functional divergence after the gene duplication. Rejection of the null hypothesis indicates that some conserved amino acid residues have experienced radical shift of amino acid property between the duplicate genes. The site-specific profile based on the posterior probabilities is useful to select residues that are type-II functional divergence related. Apparently, radical cluster-specific sites usually receive the highest scores for type-II functional divergence. However, previous computational methods have difficulty to rank ordering the conserved cluster-specific sites or imperfect radical cluster-specific sites (e.g., with one amino acid substitution within clusters) because the score system adopted was ad hoc, subject to arbitrary. Our software was developed to overcome these potential pitfalls. For instance, in the COX gene family (table 2), we found that the imperfect radical cluster-specific sites may be more important than conserved cluster-specific sites to understand the mechanism of functional divergence for this family.

As shown in tables 2 and 3, our case studies have demonstrated the critical role of amino acid classification in predicting type-II functional divergence as it determines radical or conserved change in the early stage. There have been extensive discussions about this issue (e.g., Atchley et al. 2005). As a starting point, the current system classifies 20 amino acids into 4 groups (charge positive, charge negative, hydrophilic, and hydrophobic) to characterize major types of radical amino acid substitutions, which may be sufficient for most protein families. However, the rationale of classification may be questionable in some cases. For instance, we categorize histidine (H) as charge positive. In fact, the isoelectric point of H is close to physiological pH (~7.5) and the pK_a is ~6.0. So the positive charge on H is “very” sensitive to pH changes. As a result, in some areas of the cell, H is positively charged, whereas it is not the case in other areas. Indeed, the classification of 2 groups (positive and negatively charged residues) is somewhat exaggerated as, frequently, these residues are embedded in local environment that significantly modifies the pH. Consequently, they are equivalent to serve as H-bond donor or acceptor. For instance, site 219 in table 3: this site may be incorrectly highlighted as “radical” because glycine (G) and alanine (A) are the 2 smallest amino acids, if the functional constraints that act on this site require that only small amino acids occupy this site regardless of their physiochemical properties.

Given these sophisticated cases, the software DI-VERGE2 has the option of amino acid classification for the user. It should be emphasized here that biological interpretations of type-II site predictions are based on the specified amino acid classification. One possible improvement in the future is to adopt the hierarchical amino acid classification.
for example, Murphy et al. (2000) and Li et al. (2003); the former essentially clusters PAM/BLOSUM matrix, and the latter clusters residues based on biophysics of folding. Hence, such approach allows the user to group amino acid under a given cutoff, providing a systematic approach to test whether the predicted functional residues are sensitive to amino acid grouping.

Because our method relies on the ancestral sequence inference, the accuracy of ancestral state estimations may cause some potential problems. It should be noted that the statistical method we developed for estimating $\theta_0$ and other parameters only utilized the frequencies of amino acid differences in the early stage, rather than the inferred ancestral characters. Computer simulations conducted by Zhang and Nei (1997) indicated that the frequencies of amino acid substitutions between ancestral nodes are almost unbiased, though the inferred amino acid residues at some sites may be incorrect. Moreover, in calculating the site-specific profile for prediction, we are more interested in radical/conserved cluster-specific sites or imperfect cluster-specific sites. In these cases, ancestral sequence inference is almost 100% correct. In addition, the number of amino acid substitutions in each cluster is estimated by Gu and Zhang’s (1997) algorithm, and this approach accounts for multiple hits. Further, we have used a modified U method without invoking the ancestral sequence inference: the early-stage distance $(d)$ between internal nodes $(a$ and $b)$ is approximated by the branch length estimated by the Neighbor-Joining phylogeny. In the cases of COX and G-alpha gene families, the results are very similar (not shown).

Among many statistical problems that remain to be solved in the phylogeny-based functional analysis of protein sequences, one common issue is how to model functional divergence of insertion and deletions (indels), which is important because the protein structure is more strongly affected by indels than by point substitutions. We are developing a Poisson model to integrate indel events into the framework of type-II functional divergence. This simple approach will treat indel as a 21st character, and its evolutionary rate will be indel-length $(k)$ dependent. Gu and Li (1995) found a power law for the size distribution of indels. Thus, one may assume the rate of $k$ indels can be written as $\mu_k \sim k^{-\beta}$, where $\beta \approx 2$ (Gu and Li 1995).

To accomplish the goal of phylogenomic annotation of gene functions, we need many so-called evolutionary models of protein function to link between the biochemical–physiological perspective and the evolutionary pattern of sequence. Type-I and type-II functional divergences, for instance, are 2 special models for this purpose. The statistical method we have described in this article can be integrated into a relational protein database for cluster-specific amino acid residues of, for example, transcription factors, to help in the better understanding of protein function.

Acknowledgments

The author is grateful to Jie Liang, Eric Gaucher, and Kent Vander Velden for constructive comments. This work was partially supported by the National Institutes of Health grants.

Literature Cited


Naoko Takezaki, Associate Editor

Accepted July 6, 2006