MBE Advance Access published July 24, 2006

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

Xun Gu[1]

Department of Genetics, Development and Cell Biology

Center for Bioinformatics and Biological Statistics

Iowa State University

Ames, IA 50011

Key words: Gene duplication, functional divergence

1. Address for correspondence and reprints: Xun Gu

Abstract

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 indicated that radical cluster-specific residues are responsible for most of inferred type-II functional divergence, while 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; Gaucher et al. 2002; Landgraf et al. 2001; Wang and Gu 2001; Knudson and Miyamoto 2001; Lopez et al. 2002; Jordon et 2002; Gribaldo et al 2003; Gu et al 2003; Madabushi et al. 2004; Gao et al 2005; Rastogi and Liberles 2005; Zhou et al. 2006).

Furthermore, Gu (2001) made a distinction between two types of functional divergence. Type-I functional divergence results in site-specific rate shift (Gu 1999; Gaucher et al. 2002; Landgraf et al. 2001; Knudson and Miyamoto 2001; 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, e.g., positive versus negative charge differences at a homologous site that is otherwise evolutionally conserved between subtrees within a phylogeny. Note that these two types of functional divergence may have other names. For instance, the basic Evolutionary Trace approach (Lichtarge et al. 1996; Madabush 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 as called '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; Gaucher et al. 2002; Landgraf et al. 2001; Knudson and Miyamoto 2001). While 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 two gene clusters generated by a gene duplication event, type-II functional divergence results in site-specific shift of amino acid physio-chemical 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 two 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 two 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, while in the late (L) stage, the purifying selection plays a major role to maintain related but distinct functions of two duplicate genes (Ohno 1970; Kimura 1983; Hughes 1994; Force et al. 1999). Accordingly, we modify the *two-state model* (Gu 1999; 2001) specific to type-II (cluster-specific) functional divergence:

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

(ii) In the late (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 JTT model (Jones,

Taylor, and Thornton 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 over-parameterization, we propose a simple F_1-substitution model that can distinguish between the *radical* and *conserved* amino acid substitutions. First, we tentatively classify twenty amino acids into four groups: charge positive (K, R, H), charge negative (D, E), hydrophilic (S, T, N, Q, C, G, P), and hydrophobic (A, I, L, M, F, W, V, Y). An amino acid substitution is called radical (denoted by **R**) if it changes from one group to another; otherwise it is called conserved, i.e., within the group, denoted by **C**. It should be mentioned that the conservations of residues may be due to selection pressure for ¿ 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

$$\begin{aligned} P(R|F_0) &= \pi_R(1-e^{-\lambda t}) \\ P(C|F_0) &= \pi_C(1-e^{-\lambda t}) \\ P(N|F_0) &= e^{-\lambda t} \end{aligned} \tag{1}$$

respectively, where t is the evolutionary time, λ is the substitution rate, and π_R (π_C) is the proportion of radical (or conserved) substitutions in the total substitutions; $\pi_R + \pi_C = 1$. Apparently, Eq.(1) is an extended Poisson model of protein sequence evolution. Based on the Dayhoff PAM matrix, we empirically determined $\pi_R = 0.312$ 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 $P(N|F_1) = 0$. Naturally, one may choose $P(R|F_1)$ and $P(C|F_1)$ in the same forms of Eq.(1) such as $a'_R(1 - e^{-\lambda t})$, $a'_C(1 - e^{-\lambda t})$, respectively. Together, these arguments directly lead to

$$\begin{aligned} P(R|F_1) &= a_R \\ P(C|F_1) &= a_C \\ P(N|F_1) &= 0 \end{aligned} \tag{2}$$

where $a_R = a'_R/(a'_R + a'_C)$, 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 (π_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 λ_E and λ_L be the evolutionary rates in the early (E) and late (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

$$\phi(u) = \frac{\alpha^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-\alpha u} \tag{3}$$

The shape parameter α describes the strength of rate variation among sites, that is, a small value of α 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 (λ_E) and late (λ_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 is independent of the rate component u, as indicated by Eq.(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 two clusters generated by a gene duplication event, each of which consists of several orthologous genes (figure 1). Let X be the amino acid pattern of the late stage, the column (site) of the multiple alignment of the sequences. Let $Y = (a, b)$ be the amino acid pattern of the early stage, the ancestral sequences of two 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 Eq.(1) for $Y = N, C$ or R, respectively, $P(X|Y)$ is the likelihood of the subtrees of two 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 (Felseinstein 1981; Gu 2001), and the gamma distribution density for rate variation among sites $\phi(u)$ is given by Eq.(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 Eq.(2). Remembering that the probability of a site being under F_1 is given by $P(F_1) = \theta_{II}$, the coefficient of type-II functional divergence, we have the joint probability for X and Y as follows

$$P(X, Y) = (1 - \theta_{II})P(X, Y|F_0) + \theta_{II}P(X, Y|F_1) \tag{4}$$

Direct application of Eq.(4) for estimating θ_{II} 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, e.g., 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, it 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 two 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 (k) 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, e.g., for cluster A, we have

$$p_A(k) = \frac{(\lambda_A T_A)^k}{k!} e^{-\lambda_A T_A} \tag{5}$$

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 λ_A (or λ_B) is the evolutionary rate of cluster A (or B), respectively. Hence, the early-late joint distribution can be specified as $f_{ij,Y} = P(X = (i,j), Y)$, where i or j is the number of substitutions in cluster A or B. Under this model, $P(X|Y) = p_A p_B$, which is independent of the early stage Y. Similar to the derivation of Eq.(4), we have $P(X = (i,j), Y|F_0) = \int_0^\infty P(Y|F_0) p_A(i) p_B(j) \phi(u) du$,

and $P(X=(i,j),Y|F_1) = \int_0^\infty P(Y|F_1)p_A(i)p_B(j)\phi(u)du$. Together, one can show the early-late distribution under the Poisson-based model is given by

$$f_{ij,Y} = (1-\theta_{II})\int_0^\infty P(Y|F_0)p_A(i)p_B(j)\phi(u)du + \theta_{II}a_Y\int_0^\infty p_A(i)p_B(j)\phi(u)du \quad (6)$$

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

Analytical form of the early-late distribution

First we consider the late-stage distribution, f_{ij}, the probability for i and j substitutions in clusters A and B, respectively. From Eq.(6), one can show that $f_{ij} = f_{ij,R} + f_{ij,C} + f_{ij,N} = \int_0^\infty p_A(i)p_B(j)\phi(u)du = Q_{ij}$, which is a specific version of bivariate negative binomial distribution,

$$Q_{ij} = \frac{\Gamma(i+j+\alpha)}{i!\ j!\ \Gamma(\alpha)} Z^\alpha Z_A^i Z_B^j \quad (7)$$

where $Z = \alpha/(D_A+D_B+\alpha)$, $Z_A = D_A/(D_A+D_B+\alpha)$ and $Z_B = D_B/(D_A+D_B+\alpha)$; $D_A = \bar{\lambda}_A T_A$ and $D_B = \bar{\lambda}_B T_B$ are the total branch lengths of clusters A and B, respectively, and α is the gamma shape parameter.

Next we consider the early-stage distribution f_Y, the frequencies of three early-stage amino acid patterns for $Y = N, R$ or C. Since $f_Y = \sum_{i,j} f_{ij,Y}$, from Eq.(6) one can show

$$f_Y = (1-\theta_{II})(1-e^{-d})\pi_Y + \theta_{II}a_Y, \qquad Y = R, \text{ or } C \quad (8)$$

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

$$p = (1-\theta_{II})(1-e^{-d}) + \theta_{II} \quad (9)$$

where $d = \bar{\lambda}t$ is the branch length of the early stage.

Define $W = \alpha/(D_A + D_B + d + \alpha)$, $W_A = D_A/(D_A + D_B + d + \alpha)$ 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.

$$\begin{aligned} f_{ij,N} &= (1-\theta_{II})M_{ij} \\ f_{ij,R} &= (1-\theta_{II})(Q_{ij}-M_{ij})\pi_R + \theta_E\, a_R Q_{ij} \\ f_{ij,C} &= (1-\theta_{II})(Q_{ij}-M_{ij})\pi_C + \theta_E\, a_C Q_{ij} \end{aligned} \tag{10}$$

where $M_{ij} = \int_0^\infty e^{-\lambda t} p_A(i) p_B(j) \phi(u) du$ is given by

$$M_{ij} = \frac{\Gamma(i+j+\alpha)}{i!\; j!\; \Gamma(\alpha)} W^\alpha W_A^i W_B^j \tag{11}$$

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 two clusters. As shown by Eq.(7), Q_{ij} depends on three (late-stage) parameters D_A, D_B and α. We thus modified the likelihood method of Gu and Zhang (1997) to estimate them simultaneously, denoted by $\hat{D}_A$, $\hat{D}_B$ and $\hat{\alpha}$, 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 or C. After treating three late-stage parameters as known, we develop a simple likelihood to estimate early-stage parameters θ_{II}, a_R/a_C, and d. From Eq.(10), we have $f_{ij,S} = f_{ij,R} + f_{ij,C} =$

$Q_{ij} - (1-\theta_{II})M_{ij} = Q_{ij} - f_{ij,N}$. Let $n_{ij,S} = n_{ij,R} + n_{ij,C}$. Thus, the log-likelihood function

$$\ell = \sum_{i,j} n_{ij,N} \left[\ln(1-\theta_{II}) + \ln M_{ij}\right] + \sum_{i,j} n_{ij,S} \ln(Q_{ij} - f_{ij,N}) \tag{12}$$

includes two unknown parameters θ_{II} 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 Eq.(9), the ML estimate of θ_{II} is given by $\hat{\theta} = 1/(1-y)$, where y is the solution of

$$\sum_{ij} \frac{n_{ij,S} M_{ij}}{Q_{ij} y - M_{ij}} = N_0 \tag{13}$$

with $d = -\ln(1-p) + \ln(1-\theta_{II})$. (Note that M_{ij} depends on the parameter d, while Q_{ij} only depends of late-stage 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, $\hat{f}_{ij,S} = n_{ij,S}/L$ and $\hat{f}_0 = N_0/L$. The sampling variance of $\hat{\theta}_{II}$ can be calculated as follows

$$Var(\hat{\theta}) = \frac{1}{L(\hat{f}_0 + a)} \tag{14}$$

where $a = \sum_{i,j} \hat{f}_{ij,S} M_{ij}^2 / (Q_{ij} - M_{ij} + M_{ij}\hat{\theta}_{II})^2$. When the estimates $\hat{\theta}_{II}$ and $\hat{d}$ are obtained, a_R can be estimated from Eq.(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 squared method rather than the maximum-likelihood 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 this approach for estimating p is almost unbiased.

The U-likelihood: This method utilizes amino acid sites that are universally conserved in both clusters, i.e., $i = j = 0$. Let n_{00Y} be the number of sites with $Y = N$ (the U-type), R, or C, respectively. 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

$$\ell_u = \sum_{Y=N,R,C} n_{00,Y} \ln f_{00,Y} + (N - n_{00}) \ln(1 - f_{00}) \tag{15}$$

Let $\hat{f}_{00N} = n_{00N}/N$. Similar to above, we have shown that the ML estimates of θ_{II} and d are given by

$$\begin{aligned} \theta_{II} &= 1 - \hat{f}_{00,N} \left[1 + \frac{\hat{D}_A + \hat{D}_B + d}{\hat{\alpha}} \right]^{\hat{\alpha}} \\ d &= -\ln(1-p) + \ln(1 - \theta_{II}) \end{aligned} \tag{16}$$

The sampling variance of the estimates $\hat{\theta}_{II}$ is $Var(\theta_{II}) = f_{00N}(1 - f_{00N})b^2/N$, where $b = [1 + (\hat{D}_A + \hat{D}_B + d)/\hat{\alpha}]^{\hat{\alpha}}$. Since 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 (θ_{II}) 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, i.e., $P(F_1|X, Y)$. According to the Bayesian law, we have

$$P(F_1|X, Y) = \frac{P(F_1)P(X, Y|F_1)}{P(X, Y)} \tag{17}$$

where the prior probability of F_1 in the early stage is given by $P(F_1) = \theta_{II}$. Under the Poisson-based model, $P(X = (i,j), Y|F_1)$ and $P(X = (i,j), Y|F_0)$, and $P(X = (i,j), Y)$ are given by Eqs.(5) and (7), respectively. Noting that $a_Y = 0$ if $Y = N$, one can show

$$\begin{aligned} P(F_1|X,Y) &= 0 && \text{if} \quad Y = N \\ P(F_1|X,Y) &= a_C\theta_{II}Q_{ij}/f_{ij,Y} && \text{if} \quad Y = C \\ P(F_1|X,Y) &= a_R\theta_{II}Q_{ij}/f_{ij,Y} && \text{if} \quad Y = R \end{aligned} \tag{18}$$

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

$$\begin{aligned} R(F_1|F_0) &= 0 && \text{if} \quad Y = N \\ R(F_1|F_0) &= \frac{\theta_{II}}{1-\theta_{II}}\frac{a_C}{\pi_C}\frac{1}{1-(1-h)^{i+j+\alpha}} && \text{if} \quad Y = C \\ R(F_1|F_0) &= \frac{\theta_{II}}{1-\theta_{II}}\frac{a_R}{\pi_R}\frac{1}{1-(1-h)^{i+j+\alpha}} && \text{if} \quad Y = R \end{aligned} \tag{19}$$

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

Statistical evaluation of cluster-specific sites

An important result from Eq.(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, i.e., $i = j = 0$ and $Y \neq N$. 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

$$R(F_1|F_0)_{max} = \frac{\theta_{II}}{1-\theta_{II}}\frac{a_R}{\pi_R}\frac{1}{1-(1-h)^{\alpha}} \tag{20}$$

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 θ_{II} is not significantly larger than 0. Finally, we note that $R(F_1|F_0)_{max} \to \infty$ if $h \to 0$. 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 website http://xgu.gdcb.iastate.edu. We have distribution packages for both Microsoft Windows and Linux operating systems, including manual and example files. We have conducted several case-studies to demonstrate its potential applications in understanding functional divergence of protein sequences.

Datasets: We present case-studies analyzing three gene families. (*i*) 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. There are two tissue-specific isoforms 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 1(C). (*ii*) 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 (*iii*) we have also analyzed the duplicated isoforms (G_q and G_S) 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 1(C), 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 isoforms (G_q and G_S) 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 two major CED-3 and ICE subfamilies. This raises the question whether the relative importance of type-I and type-II functional divergences is associated with specific functional classes of the protein family.

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, i.e., $R(F_1|F_0)_{max} = 7.17$. 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 radical substitutions for type II functional divergence in the early stage is

about 2.7-fold increasing ($a_R/\pi_R = 2.7$ 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, while the conserved cluster-specific sites are less likely, as indicated by a low posterior probability (~ 0.35). This case-study clearly shows the important role of statistical analysis, otherwise one cannot objectively justify whether one-less radical cluster-specific sites (i.e., there is one amino acid substitution in the late stage) is more likely to be functional divergence-related than conserved cluster-specific sites.

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). Since 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 four 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, while it is not the case in other areas. Indeed, the classification of two groups (positive and negatively charged residues) is somewhat exaggerated, as frequently these residues are embedded in local environment that significantly modifies the pK_a and make both groups of residues protonated. Consequently, they are equivalent to serve as H-bond doner or acceptor. Another example is site 419 in Table 3: It may be incorrectly highlighted this site as 'radical' because glycine (G) and alanine (A) are the two smallest amino acids, if the functional constraints that act on this site require that only small amino acids occupy this site regardless of their physio-chemical properties.

Given these sophisticated cases, the software DIVERGE2 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, e.g., Murphy, Wallqvist and Levy (2002), and Li, Hu, and Liang (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.

Since 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 θ_{II} 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 are estimated by Gu and Zhang's (1997) algorithm and that 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 NJ-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 since

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 twenty-first 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 two 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, e.g., transcription factors, to help better understand of protein function.

Acknowledgment: The author is gracefull to Jie Liang, Eric Gaucher, and Kent Vander Velden for constructive comments. This work was partially supported by the NIH grants.

Literature Cited

Atchley, W. R., Zhao, J., Fernandes, A. and Drueke, T. 2005. Solving the sequence 'metric' problem: PNAS 102: 6395-6400

Bork P, Koonin EV. 1998. Predicting functions from protein sequences – where are the bottlenecks? Nature Genetics 18:313-318.

Casari G., Sander C, Valencia. 1995. A method to predict functional residues in proteins. Structural biology 2:171-178.

Chothia C., Lesk A. M. 1996. The relation between the divergence of sequence and structure in proteins. EMBO J 5:823-826.

Eisen, M. B and Fraser, C. M. 2003. Phylogenomics: intersection of evolution and genomics. Science 300:1706-1707.

Gao X., K. Vander Velden, D. F. Voytas, and X. Gu. 2005. SplitTester: software to identify domains responsible for functional divergence in protein family. BMC Bioinformatics, 6:137

Gaucher EA, Gu X, Miyamoto, M, Benner, S. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. Trend in Biochemical Sciences (TiBS) 27:315-321.

Golding GB, Dean AM. 1998. The structural basis of molecular adaptation. Mol Biol Evol 15:355-369

Gribaldo, S. D. Casane, P. Lopez,a and H Philippea. 2003. Functional divergence prediction from evolutionary analysis: A case Study of Vertebrate Hemoglobin. Mol. Bio. Evol. 20:1754-1759.

Gu, J, Wang Y, Gu X. 2002. Pattern of Functional divergence in JAK tyrosine protein kinase family. J. Mol. Evol. 54:725-733

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

Gu, X. 2001. Maximum likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. 18: 453-464.

Gu, X. 2003. Functional divergence in protein (family) sequence evolution. Genetica 118:133-141.

Gu, X., K. Vander Velden. 2002. DIVERGE: Phylogeny-based Analysis for Functional-Structural Divergence of a Protein Family. Bioinformatics 18:500-501.

Jordan, I.K. et al. 2001. Sequence and structural aspects of functional diversification in class I a-mannosidase evolution. Bioinformatics 17, 965-976

Knudsen, B. and Miyamoto, M.M. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. PNAS 98, 14512-14517.

Landgraf R., I. Xenarios, and D. Eisenberg. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J. Mol. Biol. 307:1487-1502.

Li, X., C. Hu, and J. Liang. 2003. Simplicial edge representation of protein structures and alpha contact potential with confidence measure. Proteins. 53:792-805.

Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257:342-358.

Lopez P., P. Forterre, and H. Philippe. 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. 49(4):496-508.

Madabushi, S., Gross, A.K., Philippi, A., Meng, E.C., Wensel, T.G., and Lichtarge, O. 2004. Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. J. Biol. Chem. 279: 8126?132.

Murphy L, Wallqvist A, Levy R. 2000. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein 13:149-152.

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.

Rastogi, S and Liberles, D. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evolutionary Biology 2005, 5:28

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

Thorne JL, Goldman N, Jones DT. 1996. Combining protein evolution and secondary structure. Mol Biol Evol 13:666-673.

Tseng Y. Y. and J. Liang. 2006. Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: A Bayesian Monte Carlo approach Mol. Biol. Evo. 23:421-436.

Wang Y, Gu X. 2001. Predicting functional divergence of caspase gene family. Genetics. 158:1311-1320.

Yang, Z., S. Kumar, and M. Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics. 141:1641-1650.

Zhang J, Nei, M. 1997. Accuracies of Ancestral amino acid sequence inferred by the Parsimony, Likelihood, and distance method. J. Mol. E. 44 (suppl 1): S139-S146

H Zhou, J Gu, S J. Lamont, Gu X. 2006 Evolutionary analysis for functional divergence of Toll-like receptor gene family and altered functional constraints. J. Mol Evol (in press)

Figure Legends:

Figure 1. (A) Two gene clusters after gene duplication. E and L are early and late stages of gene cluster A and B, respectively. (B) Type I functional divergences: in the early stage , the evolutionary rate may increase for functional divergence-related change, resulting in shifted functional constraints between clusters A and B. Type II functional divergence: in the early stage , the evolutionary rate may increase for functional divergence-related change, resulting in radical shift in amino acid property, but in the late stage is back to the same level of sequence conservation. (C) The phylogenetic tree of COX gene family, which was inferred by the neighbor-joining method, using amino acid sequences with Poisson distance. Bootstrapping values more than 50% are presented. Modified from Gu (1999) and Gu (2001).

Figure 2. Site-Specific profile for type-II functional divergence between COX-1 and COX-2, measured by the posterior ratio. Horizontal lines (1)-(4) indicate cluster-specific patterns in Table 2.

Table 1. Summary of functional divergence analysis for COX and G-protein alpha families.

Type II	*COX*	*G-protein alpha*
N	370	151
C	102	72
R	111	111
p	0.365	0.548
D_a	0.376	0.820
D_b	0.590	0.944
d	0.282	0.402
α	0.401	0.440
f_R	0.521	0.607
a_R/π_R	2.744	2.811
θ_{II} + *s.e.*	0.159±0.036	0.325±0.055
Type I		
θ_I + *s.e.*	0.490±0.085	0.436±0.071

Note: *N*, *C* and *R* are the numbers of sites across internal nodes (*a*, *b*) of the tree (see Figure 1, panel *A*) that display no difference, conserved difference and radical differences, respectively, and *p* is the proportion of (overall) differences between nodes *a* and *b*. D_a and D_a 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 α is the gamma shape parameter. f_R is the observed proportion of radical changes in all substitutions between nodes *a* and *b*. a_R/π_R is the ratio of radical changes under (type II) functional divergence versus non functional divergence. Finally, θ_I *and* θ_{II} are the coefficients of type I or type II functional divergence, respectively.

Table 2. Functional Ranking of several cluster-specific patterns in the COX gene family

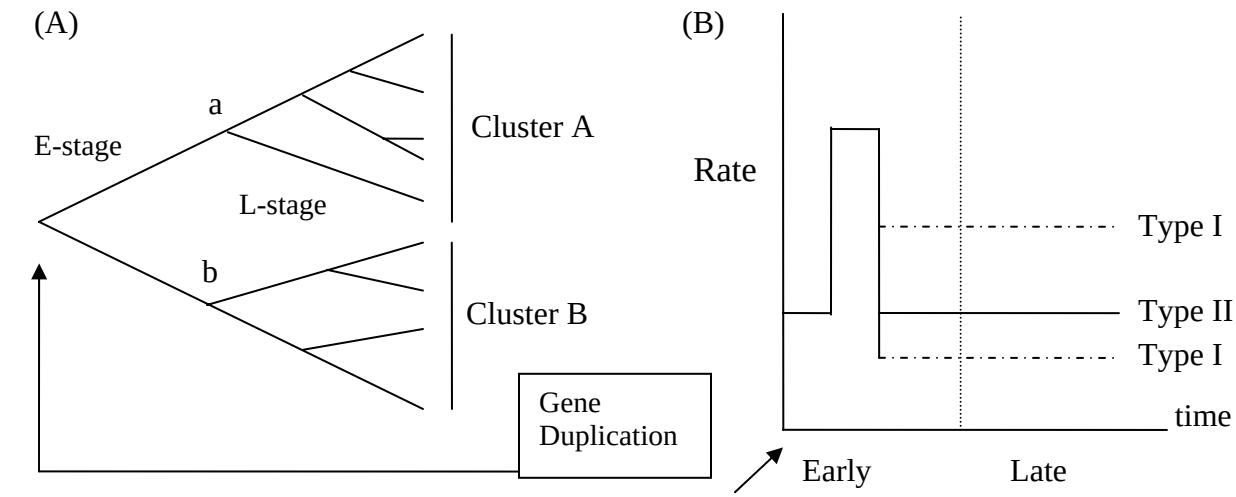
	Between clusters (Early-stage)	*Within clusters (Late-stage)*	*Num. of Sites*	*Ratio score*	*Posterior Prob.*
(1)	Radical change (Radical cluster-specific)	No a.a. change	28	7.17	0.88
(2)	Radical change	One a. a. change	30	2.11-2.22	0.68-0.69
(3)	Radical change	Two a.a. changes	20	1.25-1.41	0.56-0.59
	…	…	…	…	…
(4)	Conserved change (Conserved cluster-specific)	No a.a change	31	0.55	0.35

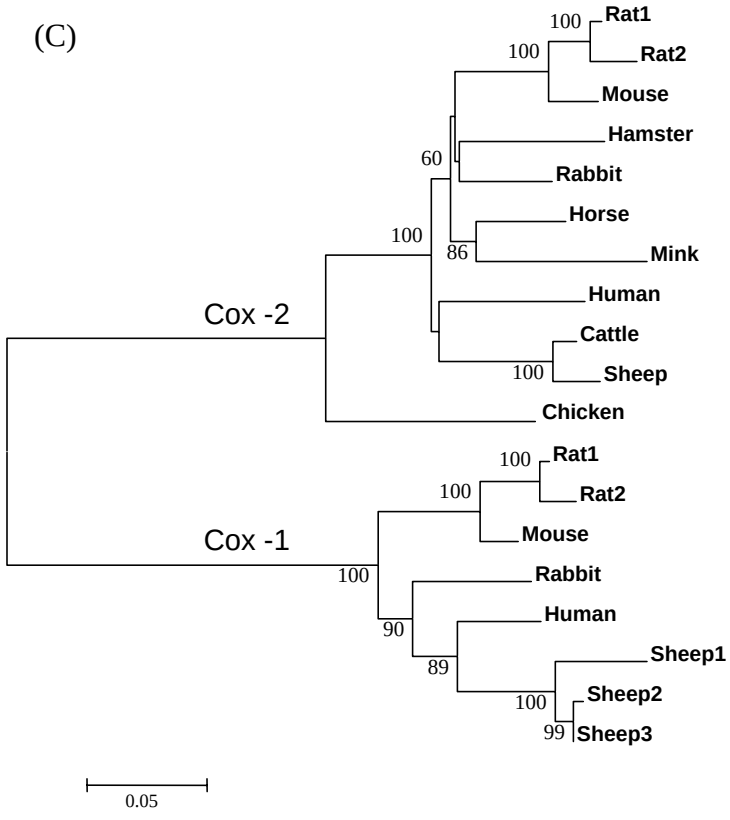
Pattern (1): Radical cluster-specific sites. Patterns (2)-(3): Imperfect radical cluster-specific sites. Pattern (4): Conserved cluster-specific sites.

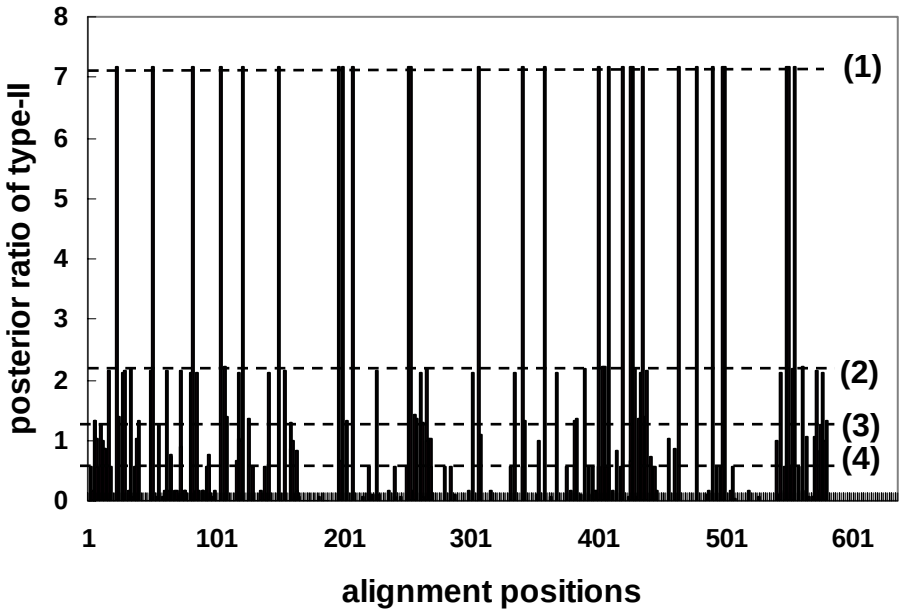
Table 3. Summary of amino acids changes in 22 radical cluster-specific positions associated with the divergence of COX1 and COX2.

Position	ClOX1	COX2	Property change
22	Y	S	H vs P0
51	P	E	P0 vs -
82	W	G	H vs P0
103	V	S	H vs P0
121	I	K	H vs +
149	T	V	P0 vs H
197	S	D	P0 vs -
251	E	K	- vs +
253	A	T	H vs P0
306	T	E	P0 vs -
340	F	H	H vs +
358	R	Q	+ vs P0
401	Y	H	H vs +
409	A	S	H vs P0
419	G	A	P0 vs H
425	D	P	- vs P0
427	H	A	+ vs H
435	V	S	H vs P0
463	Q	E	P0 vs -
499	S	A	P0 vs H
548	K	Q	+ vs P0
555	T	V	P0 vs H

H: hydrophobic; P0: hydrophilic with neutral charge; +: charge positive; and -: charge negative.


(A)
E-stage
a
L-stage
b
Cluster A
Cluster B
Gene Duplication
(B)
Rate
Type I
Type II
Type I
time
Early
Late


(C)
Rat1
Rat2
Mouse
Hamster
Rabbit
Horse
Mink
Human
Cattle
Sheep
Chicken
Cox -2
Rat1
Rat2
Mouse
Rabbit
Human
Sheep1
Sheep2
Sheep3
Cox -1
100
100
60
100
86
100
100
100
100
90
89
100
99
0.05


posterior ratio of type-II
8
7
6
5
4
3
2
1
0
(1)
(2)
(3)
(4)
1
101
201
301
401
501
601
alignment positions