An Update of DIVERGE Software for Functional Divergence Analysis of Protein Family

Xun Gu,*[,1,2] Yangyun Zou,[1] Zhixi Su,[1] Wei Huang,[1] Zhan Zhou,[1] Zebulun Arendsee,[2] and Yanwu Zeng[3]

[1]State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China

[2]Department of Genetics, Development and Cell Biology, Program of Bioinformatics and Computational Biology, Iowa State University

[3]Shanghai Stem Cell Institute, Institutes of Medical Sciences, Shanghai Jiao Tong University School of Medicine, Shanghai, China

***Corresponding author:** E-mail: xgu@iastate.edu.

Associate editor: Sudhir Kumar

Abstract

DIVERGE is a software system for phylogeny-based analyses of protein family evolution and functional divergence. It provides a suite of statistical tools for selection and prioritization of the amino acid sites that are responsible for the functional divergence of a gene family. The synergistic efforts of DIVERGE and other methods have convincingly demonstrated that the pattern of rate change at a particular amino acid site may contain insightful information about the underlying functional divergence following gene duplication. These predicted sites may be used as candidates for further experiments. We are now releasing an updated version of DIVERGE with the following improvements: 1) a feasible approach to examining functional divergence in nearly complete sequences by including deletions and insertions (indels); 2) the calculation of the false discovery rate of functionally diverging sites; 3) estimation of the effective number of functional divergence-related sites that is reliable and insensitive to cutoffs; 4) a statistical test for asymmetric functional divergence; and 5) a new method to infer functional divergence specific to a given duplicate cluster. In addition, we have made efforts to improve software design and produce a well-written software manual for the general user.

Key words: **DIVERGE, type I functional divergence, type II functional divergence, gene duplication, gene family evolution.**

Letter

Because gene duplications provide raw materials for functional innovations (Ohno 1970; Wolfe and Shields 1997; Gu et al. 2002b; Eisen and Fraser 2003; Wang et al. 2009; Zou et al. 2011), it is desirable to develop sequence-based methods to identify amino acid sites that are likely responsible for any functional changes (Casari et al. 1995; Tatusov et al. 1997; Gaucher et al. 2002; Gu 2003). To facilitate such predictions, we developed software packages DIVERGE1.0 and DIVERGE2.0 (available at http://www.xungulab.com). The DIVERGE packages have been widely used (Wang and Gu 2001; Gu et al. 2002a, 2002b; Zheng et al. 2007; Zhou et al. 2007; Mertz et al. 2009), as recently reviewed by Benitez-Paez et al. (2012). DIVERGE provides a suite of statistical tools for the selection and prioritization of functional divergence-related amino acid sites within a gene family based on functional constraint patterns extracted from a multiple sequence alignment (MSA) (Gu 1999, 2001a, 2006; Gu and Vander Velden 2002).

Many other computational methods have been developed in the last decade for the same purpose. Abhiman and Sonnhammer (2005a, 2005b) developed a large-scale database (FunShift) for analyzing rate shifts between gene subfamilies, Donald and Shakhnovich (2009) for database SDR, and Lopez et al. (2007) for database FirDB. Meanwhile, a number of tools have been developed to quickly survey the functional divergence of large gene families (Kalinina et al. 2004; Pazos and Sternberg 2004; Arnau et al. 2006; Pazos et al. 2006; Brandt et al. 2010). Also, advanced multivariate analysis techniques have been used to statistically identify the determinants of funtional divergence (Casari et al. 1995; Reva et al. 2007; Wallace and Higgins 2007; Capra and Singh 2008; Bharatham et al. 2011). In many studies, protein structure information has been incorporated into the algorithm for functional divergence prediction (Lichtarge et al. 1996; Hannenhalli and Russell 2000; Landgraf et al. 2001; Blouin et al. 2003; Kalinina et al. 2004; Chakrabarti et al. 2007; Chakrabarti and Panchenko 2009; Rausell et al. 2010). Additionally, Huang and Golding (2011) and Gao et al. (2005) have developed practical methods to infer sequence regions that are under functional divergence between duplicate genes. A number of studies have shown that modeling functional divergence is closely related to a special class of evolutionary models called covarion or heterotachy (Knudsen and Miyamoto 2001; Lopez et al. 2002; Pupko and Galtier 2002; Susko et al. 2002). These efforts have demonstrated that substantial information about site-specific functional divergence can be inferred from the pattern of rate change in an MSA. Possibly coupled with protein structure and/or functional genomics data, these predicted sites can be used as candidates for further experimentation.

Diverge has several advantages over other programs, including an explicit evolutionary model and statistically

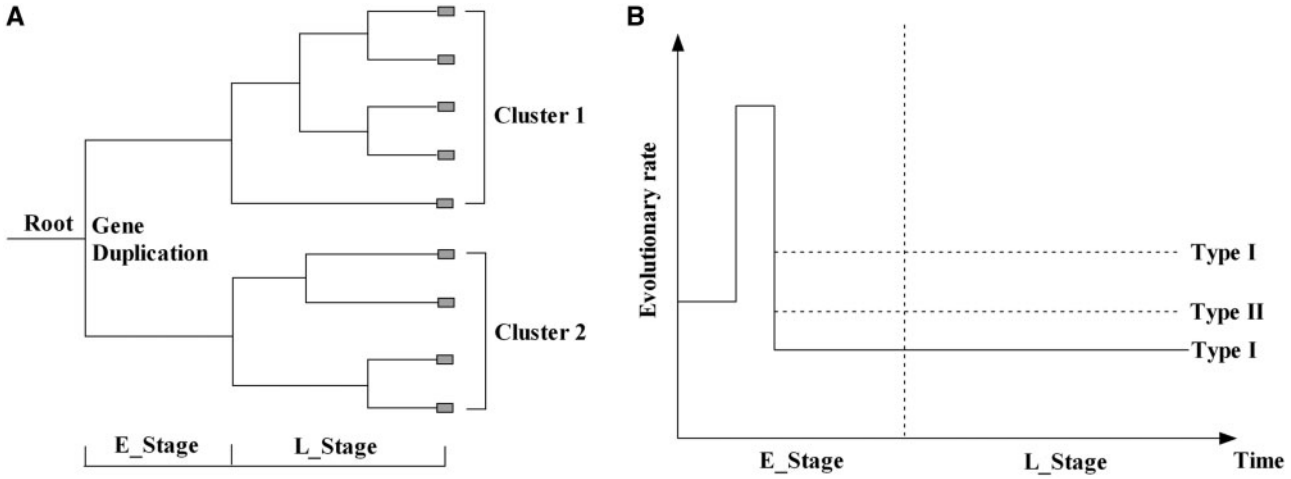
rigorous scores for site predictions. However, DIVERGE (and other programs) cannot perfectly distinguish between neutral and adaptive evolutionary changes. For instance, apparent sequence-level functional divergence between paralogous genes can be adaptive due to change in functionality or can be neutral due to biased mutational processes when the duplicated gene is located in a GC-poor isochore, whereas the original is in a GC-rich isochore. Apparently, a comprehensive analysis pipeline integrating coding and noncoding sequences with multiple functional genomics data is needed.

In response to popular demand, here we report the release of an updated version, DIVERGE3. Available at the website http://www.xungulab.com, this new version has the following features: implementation of new analytical methods, user-inspired improvements in software, and a well-written software manual for the general users.

New Methods

Statistically predicting amino acid sites that are involved in functional divergence of a protein family under the background of neutral evolution (Kimura 1983) is the central purpose of DIVERGE. In molecular evolution, the functional importance of a gene is quantified as the intensity of purifying selection (functional constraint). Because a (negatively) larger value of the selection intensity means greater functional constraint and lower evolutionary rate, changes in evolutionary rate can be interpreted as "changes in functional constraints," an indication of functional divergence. In the following, we use a simple case of two duplicate clusters (fig. 1) to illustrate how we can use DIVERGE to analyze functional divergence based on protein sequence.

Given the MSA of a gene family with two duplicate clusters, amino acid sites can be grouped into four types. 1) Type 0 represents amino acid patterns that are universally conserved throughout the whole gene family; these sites are important for the common protein structure and function. 2) Type I represents amino acid patterns that are highly conserved in one duplicate cluster but highly variable in the other; these sites may have experienced shifted functional constraints. 3) Type II represents amino acid patterns that are



C

	Sequences	Type 0	Type I	Type II	Type U
Gene I	1	RGD	DGNR	GHI	SKQA
	2	RGD	DGKH	GHI	SKQA
	3	RGD	DGNR	GHI	SRLA
	4	RGD	DGNS	GHI	SRLA
	5	RGD	DGKQ	GHI	SKLS
	6	RGD	DGNQ	GHI	PKLS
	7	RGD	DGSP	GHI	AAER
	8	RGD	DGYN	GHI	KMLK
Gene II	9	RGD	DDEL	DPG	STKT
	10	RGD	HKEL	DPG	RTKT
	11	RGD	ASEL	DPG	IKKT
	12	RGD	NSEL	DPG	KTKS
	13	RGD	QSEL	DPG	PVEK
	14	RGD	DSEL	DPG	PVEK
	15	RGD	ESEL	DPG	PTEE
	16	RGD	ESEL	DPG	PTEE
	17	RND	DREL	DPG	GKKD

FIG. 1. (*A*) Two gene clusters after gene duplication. E and L are early and late stages of gene clusters 1 and 2, respectively. (*B*) Type I and type II functional divergences after gene duplication. In the early stage, the evolutionary rate (say, in cluster 1) may increase for functional divergence-related change, but in the late stage, it may be higher (or lower) than its original rate, resulting in shifted functional constraints between clusters 1 and 2, or type I functional divergence. If the rate in the late stage is back to the same as the original one, no shifted functional constraints between clusters 1 and 2 can be observed, or type II functional divergence. (*C*) A hypothetical multiple alignment to show universally conserved sites (type 0), type I and type II amino acid patterns, and U-type sites (unclassified). Figure modified from Gu (1999, 2001a, 2006).

highly conserved within both duplicate clusters but are conserved in a biochemically different state. For example, negatively charged amino acids may be conserved in one gene and positively charged ones in the other. These sites may be involved in functional specifications. Finally, 4) Type U represents amino acid patterns that cannot be classified into the three above types, for instance, amino acid sites that are highly variable in both clusters.

The goal of DIVERGE is to develop statistical methods to test and predict type I and type II amino acid patterns. Type I functional divergence postulates that functional divergence between duplicate genes results in shifted functional constraints (i.e., different evolutionary rate) at some sites, called F_1 site, which typically show type I amino acid patterns. A two-state model was implemented. Let λ_1 and λ_2 be the evolutionary rates in duplicate clusters 1 and 2, respectively. Under the functional divergence-unrelated state (F_0), the evolutionary rate is similar between duplicate clusters ($\lambda_1 = \lambda_2$). In contrast, under the type I functional divergence-related state (F_1), λ_1 and λ_2 are independent as the result of shifted functional constraints. The key step is then to estimate the coefficient of type I functional divergence, θ_I, defined as the probability of a site being in the state F_1. Moreover, DIVERGE provided a site-specific profile based on posterior scores to predict the sites responsible for type I functional divergence.

Type II functional divergence (Gu 2006) postulates that functional divergence between duplicate genes at some sites results in type II amino acid patterns. At a site related to type II functional divergence, called F_2 site, rapid evolution occurred in the early stage after gene duplication, resulting in a radical amino acid change. Subsequently, the site became highly conserved. Here, we will briefly explain the principles behind the analysis. Although type II functional divergence may have occurred in the early (E) stage after the gene duplication, purifying selection plays a major role in the late (L) stage to maintain the related, but distinct, functions of duplicate genes. A two-state model was implemented to distinguish between them: 1) in the E stage, an amino acid site has two states: F_0 (functional divergence unrelated) and F_2 (type II functional divergence related). The probability of a site being under F_2 is given by $P(F_2) = \theta_{II}$, the coefficient of type II functional divergence. 2) In the late (L) stage, an amino acid site is always under the state of F_0, indicating that amino acid substitutions in this stage are mainly under the purifying selection. Next, let λ_E and λ_L be the evolutionary rates in the early (E) and late (L) stages, respectively. The relationship between them depends on the status of type II functional divergence. Under F_0, the evolutionary rate at a site remains the same between the early (λ_E) and late (λ_L) stages ($\lambda_E = \lambda_L$). Under F_2, λ_L and λ_E are independent. DIVERGE2.0 formulated a statistical method for the estimation of θ_{II} and implemented site-specific posterior profiles for predicting type II functional divergence.

Involvement of Gaps

All methods implemented in previous DIVERGE versions followed the common practice in phylogenetic analysis of ignoring indels (i.e., deletions and insertions). Consequently, some important functional information may be missed. A simple approach to examining functional divergence in nearly complete sequences is to represent a gap by adding an additional character (Edwards and Shields 2004). Two drawbacks of this approach are overcounting of deletion–insertion events and sensitivity to the alignment uncertainty. For instance, adjacent indels would be considered two independent events, rather than the more biologically reasonable single, two-site indel event (Gu and Li 1995). The new version of DIVERGE has implemented the following algorithm to deal with gaps. 1) Carry out functional divergence analyses, for example, estimation of the coefficients of type I and type II functional divergences, based on the alignment positions without any gap. 2) Calculate the site-specific posterior profile for those sites without gaps. 3) Given the parameters estimated for ungapped sites, calculate the site-specific posterior probability for sites with gaps by treating indels as additional characters. Our approach attempts to both avoid the effects of alignment uncertainty on phylogenetic analysis and expand the site-specific posterior profile to include as many sites as possible.

False Discovery Rate for Predicted Amino Acid Sites

Knowing the false discovery rate (FDR) of the predicted sites is critical to assessing the reliability of the results. In general, FDR is the proportion of predicted sites that are actually unrelated to functional divergence. DIVERGE and DIVERGE2 mainly use a site-specific posterior profile, denoted by Q_k for site k, as a scoring system to identify functional divergence-related amino acids. We calculate FDR with the following procedure in the updated version of DIVERGE. Let L_c be the number of sites predicted under the posterior cutoff c. Then, we have shown (Gu 2011) that FDR(c) can be approximately calculated by

$$\mathrm{FDR}(c) = 1 - \sum_{k \text{ in } A} Q_k / L_c,$$

where set A is for all sites k that satisfy $Q_k > c$. This value may help to evaluate the cost of experiments caused by false-positive predictions. There are several versions of FDR, as numerically illustrated in Xia (2011). We shall study these measures and, if we find them to be useful, implement them in the following updated version.

Effective Number of Sites Involved in Functional Divergence

Even though most studies pointed to a small number of sites that can be predicted as type I or type II functional divergence-related (Abhiman and Sonnhammer 2005a, 2005b), calculation of the average percentage of amino acid sites involved is problematic. From our preliminary analysis, we notice that, after removing those predicted sites with the strongest signals, the functional divergence between duplicate genes for the rest of amino acid sites usually becomes trivial. On the basis of this observation, we have designed a rapid nonparametric procedure to count the effective number of functional divergence-related sites.

Consider a gene family with two duplicate clusters. The effective number (n_e) of functional divergence-related sites (F sites) is defined as the minimum number of sites, such that, when they are removed, the coefficient of functional divergence for the rest of sites approaches to zero. The algorithm is as follows. 1) Obtain the site-specific posterior profile for type I or type II functional divergence by conventional methods. 2) Rank amino acid sites according to this profile. 3) Calculate the coefficient of functional divergence (denoted by θ*) after removing the site with the highest posterior probability. 4) Repeat step 3 sequentially as long as the condition $\theta^* > \text{se}^*$ holds, where se* is the standard error of θ* used to control the long-tail problem (i.e., the broadening of the null distribution when the sample size is small). 5) Stop the procedure when $\theta^* < \text{se}^*$. 6) Finally, count the number of removed sites as the number of effective F sites (n_e).

Site-Specific Rate Change After Gene Duplication

In type I functional divergence analysis, it is useful to calculate the magnitude of the site-specific rate shift between two duplicate genes. Let $r_{k,1}$ (or $r_{k,2}$) be the relative evolutionary rate at site k in duplicate cluster 1 (or cluster 2). We implemented the method of Gu (2001b) for calculating $r_{k,1}$ and $r_{k,2}$ because it was designed specifically for type I functional divergence. However, other methods would give similar results (not shown). The overall level of functional constraint of each duplicate gene can be measured by the conventional $\omega = d_N/d_S$ ratio, which may differ between the genes, that is, $\omega_1 \neq \omega_2$. We thus used the formulas $\omega_{k,1} = \omega_1 r_{k,1}$ and $\omega_{k,2} = \omega_2 r_{k,2}$ as proxies for the site-specific measurement of functional constraint. In the updated software, the change ratio of functional constraints at a site between duplicate clusters is plotted against the positions of protein sequence alignment.

Testing Asymmetric Type I Functional Divergence

DIVERGE2 implemented functional distance analysis to demonstrate the asymmetry of type I functional divergence but lacked a rigorous statistical basis. Here, we solve this problem by implementing a simple method as follows. Suppose we test whether type I functional divergence is asymmetric between duplicate clusters 1 and 2, given a more ancient duplicate cluster 3 as outgroup. Let θ_{12}, θ_{13}, and θ_{23} be the coefficients of type I functional divergence between pair-wise duplicate clusters. Under the hypothesis of symmetry between duplicate clusters 1 and 2, we have the null $\theta_{12} = \theta_{13}$ and develop an approximate method to calculate the sampling variance of $\delta = \theta_{13} - \theta_{23}$, for testing whether the null hypothesis $\delta = 0$ can be statistically rejected (see DIVERGE3 software manual for technical details).

Site-Specific Posterior Profile of Gene-Specific Type I Functional Divergence

The updated DIVERGE implements a new method that helps the user infer type I functional divergence specific to a given duplicate cluster. To this end, we need to analyze three duplicate clusters simultaneously. Under the two-state model (functional divergence unrelated F_0 or related F_1) (Gu 1999, 2001a), there are eight possible combined states for three duplicate clusters, which can be reduced to five nondegenerate patterns. $S_0 = (F_0, F_0, F_0)$ means no type I divergence occurred in any clusters. $S_1 = (F_1, F_0, F_0)$ means type I functional divergence occurred only in cluster 1, and similarly $S_2 = (F_0, F_1, F_0)$ and $S_3 = (F_0, F_0, F_1)$. The final pattern S_4 is for the rest of four states, each of which has two or three clusters that have experienced type I functional divergence.

Let $f_k = P(S_k)$, $k = 0, \ldots, 4$, be the probability of the kth (nondegenerate) pattern. We then claim that the coefficient of type I functional divergence between any two clusters is given by $\theta_{12} = f_1 + f_2 + f_4$, $\theta_{13} = f_1 + f_3 + f_4$, or $\theta_{23} = f_2 + f_3 + f_4$, respectively. This is because, say, θ_{12} includes the probabilities of type I functional divergence occurred in cluster 1 (f_1), or cluster 2 (f_2), or both (f_4). Our goal is to calculate the posterior probability of the jth joint pattern S_j conditional on the observations **x.** By Bayes rule, this is

$$P(S_j \mid \mathbf{x}) = f_j P(\mathbf{x} \mid S_j)/P(\mathbf{x}),$$

where $j = 0, 1, 2, 3$, and 4. This formula can be used to predict amino acid sites that have experienced type I functional divergence in a specific duplicate cluster. The updated version of DIVERGE implements this statistical method for estimating f_k and $P(\mathbf{x} \mid S_j)$ under the model of type I functional divergence.

Results and Discussion

We have implemented these newly developed methods in the updated version DIVERGE3 (table 1) and carried out substantial case studies to demonstrate the usefulness of these methods. We also conducted some comparative studies with those that have been addressed experimentally. Because of the space limitation of a software-type article, here we will only discuss three examples.

COX (Cyclooxygenase) Duplicate Genes (COX-1 and COX-2)

FDR for Predicted Type I Functional Divergence-Related Sites

Given the gene phylogeny and 584 aligned amino acid sites, we obtained $\theta_I = 0.56 \pm 0.11$ between COX-1 and COX-2 (Gu 2001a). For the top five predicted sites under the posterior cutoff 0.80, we calculated the corresponding FDR as 11.0%. If we lower the posterior cutoff to 0.7, the FDR value for the 19 predicted sites is 20.1%. This example shows that in practice, one can also use FDR as a primary criterion to make functional predictions.

Effective Number of Functional Divergence-Related Sites

Figure 2 shows the θ*-RemovedSite profile of type I functional divergence between COX1 and COX2. As expected, the θ* value decreases rapidly when more sites are removed and is nearly zero when the top 30 sites are removed (roughly at the posterior probability of 0.63). Because this profile has a long tail around zero, we recommend the use of one standard error of θ* to control the long tail problem; in this case, we obtain $n_e = 27$. We thus conclude that about 4.6% of amino acid sites may have been involved in the type I functional divergence between COX1 and COX2.

Table 1. A Brief Description for Analysis Options Implemented in the Updated Version, DIVERGE3.

Function	Description
Gu99	Detect type I functional divergence by Gu (1999) method
Gu2001	Detect type I functional divergence by Gu (2001a) method
Type II divergence	Detect type II functional divergence of gene family
Rate variation among sites (RVS)	Estimate the among-site rate variations for given cluster as described in Gu and Zhang (1997)
Ancestral sequence inference	Infer the ancestral sequence for each internal node
Functional distance analysis	Estimate the type I functional distance for each pair of clusters and show the type I functional branch length of each cluster when at least three homologous gene clusters are available
Newly implemented	
Gap involvement	Site-specific posterior profile for sites containing gaps
FDR for predictions	Provides more statistical evaluations for predicted sites
Asymmetric test for type I functional divergence	Statistically testing whether the degree of type I functional divergence differs between two duplicate genes
Effective number of sites related to functional divergence (type I or type II)	Estimate effectively the number of sites related to type I and type II functional divergences , which is insensitive to the cutoff
Gene-specific type I analysis	Site-specific posterior profiles for predicting gene-specific type I functional divergence-related sites

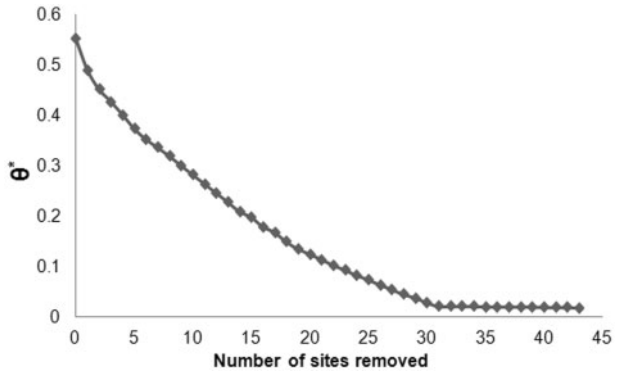


Fig. 2. The θ*-RemovedSite profile of type I functional divergence between COX1 and COX2. The θ* value decreases rapidly when more sites are removed and virtually approaches to the turning point when the top 30 sites are removed. Since then one can observe a long tail of this profile around zero.

Site-Specific Rate Change between COX1 and COX2

We calculated the site-specific $\omega = d_N/d_S$ ratio in COX1 and COX2 duplicate clusters (the mean $\omega = d_N/d_S$ for COX1 is 0.076/0.707 = 0.108 and that for COX2 is 0.076/0.551 = 0.137). The change in site-specific rate ranges approximately from 4- to 8-fold at predicted type I functional divergence-related sites (as shown in supplementary fig. S2, Supplementary Material online).

Vertebrate Developmental Gene Families

Testing Asymmetric Type I Functional Divergence

We have applied the asymmetry test to 10 vertebrate developmental gene families (supplementary fig. S1, Supplementary Material online). The results are summarized in supplementary table S1, Supplementary Material online. Four gene families show a highly significant asymmetric pattern of type I functional divergence ($P < 0.005$), two are significantly asymmetric ($P < 0.05$ or $P < 0.01$), and the other four are symmetric.

Site-Specific Posterior Profile of Gene-Specific Type I Divergence in ADRA2

The alpha2-adrenergic receptors (ADRA2) gene family in vertebrates usually contains three duplicate clusters: ADRA2A, ADRA2B, and ADRA2C. The phylogeny of ADRA2 family (supplementary fig. S3, Supplementary Material online) indicates that these duplicate genes may have been generated in the early stage of vertebrate evolution. Figure 3 shows site-specific profiles of ADRA2A, ADRA2B, and ADRA2C-specific type I divergence in panels *A*, *B*, and *C*, respectively. Given ADRA2A cluster as the outgroup (supplementary fig. S3, Supplementary Material online), we observed seven type I sites predicted in ADRA2B (positions 120, 190, 245, 247 and 350, 351, and 421) under the posterior cutoff 0.6, whereas there is only one predicted type I site (position 258) in ADRA2C. This observation is consistent with the result of the asymmetric test for type I functional divergence between ADRA2B and ADRA2C ($P < 0.005$).

Outlook

Many factors can affect divergence in protein evolution, such as genomic bias of GC content (Su et al. 2011), histone modifications (Zou et al. 2012), alternative splicing (Su et al. 2006; Su and Gu 2012), and tissue expression pattern (Huang et al. 2009). In other words, amino acid substitution is just one of many evolutionary mechanisms. Therefore, combining computational analyses and experimental data sets is the current trend to solve the central problem in protein evolution. This synergistic effort permits a more rational, cost-effective design of functional analysis of proteins. DIVERGE has been widely used for this purpose. DIVERGE3 will have the flexibility to be integrated into computational pipelines to ease the analysis of increasingly large data sets, with the option of open source.

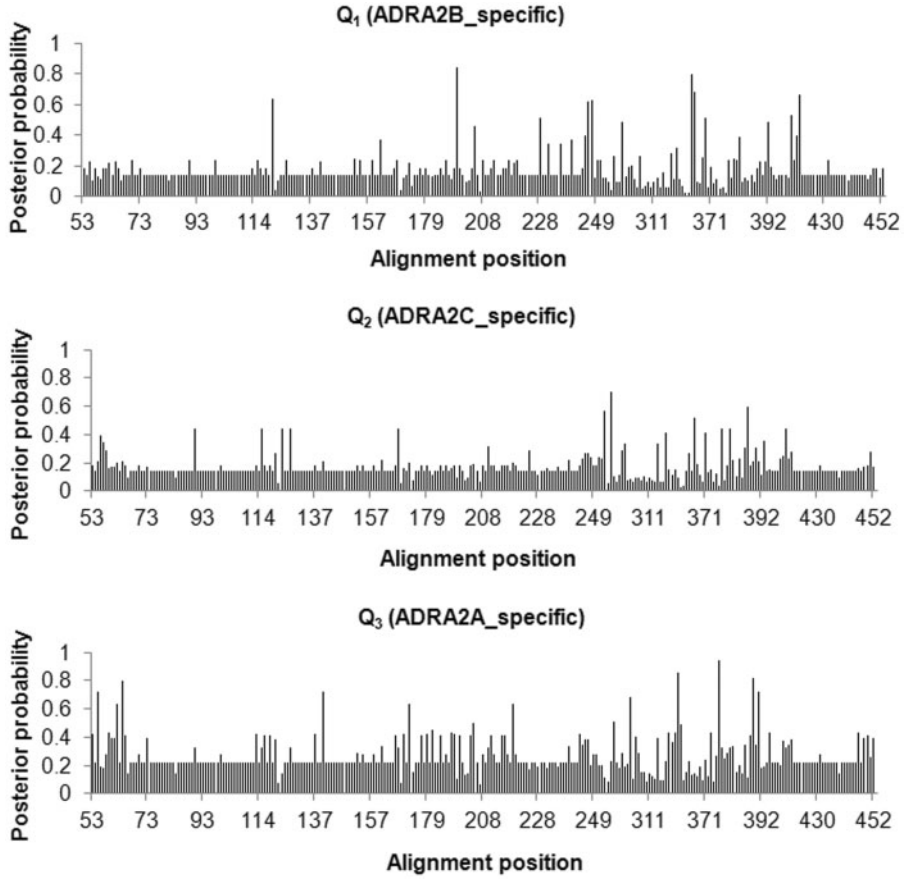


FIG. 3. Site-specific profile for predicting critical amino acid residues responsible for the gene-specific type I functional divergence among ADRA2 subfamily, measured by posterior probability. Site-specific profile of amino acid sites responsible for the (*A*) ADRA2B, (*B*) ADRA2C, and (*C*) ADRA2A-specific type I functional divergence.

Supplementary Material

Supplementary figures S1–S3 and table S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to the reviewing editor and two anonymous reviewers for their constructive comments in the early version of manuscript. This work was partly supported by grants from Fudan University and Iowa State University, and a grant from the Ministry of Science and Technology China (2012CB910101). Y.Z. was supported by Specialized Research Fund for the Doctoral Program of Higher Education of China (New Teachers, NO. 20120071120009).

References

Abhiman S, Sonnhammer EL. 2005a. FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res.* 33:D197–D200.

Abhiman S, Sonnhammer EL. 2005b. Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins* 60:758–768.

Arnau V, Gallach M, Lucas JI, Marin I. 2006. UVPAR: fast detection of functional shifts in duplicate genes. *BMC Bioinformatics* 7:174.

Benitez-Paez A, Cardenas-Brito S, Gutierrez AJ. 2012. A practical guide for the computational selection of residues to be experimentally characterized in protein families. *Brief Bioinform.* 13:329–336.

Bharatham K, Zhang ZH, Mihalek I. 2011. Determinants, discriminants, conserved residues—a heuristic approach to detection of functional divergence in protein families. *PLoS One* 6:e24382.

Blouin C, Boucher Y, Roger AJ. 2003. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. *Nucleic Acids Res.* 31:790–797.

Brandt BW, Feenstra KA, Heringa J. 2010. Multi-harmony: detecting functional specificity from sequence alignment. *Nucleic Acids Res.* 38:W35–W40.

Capra JA, Singh M. 2008. Characterization and prediction of residues determining protein functional specificity. *Bioinformatics* 24:1473–1480.

Casari G, Sander C, Valencia A. 1995. A method to predict functional residues in proteins. *Nat Struct Biol.* 2:171–178.

Chakrabarti S, Bryant SH, Panchenko AR. 2007. Functional specificity lies within the properties and evolutionary changes of amino acids. *J Mol Biol.* 373:801–810.

Chakrabarti S, Panchenko AR. 2009. Ensemble approach to predict specificity determinants: benchmarking and validation. *BMC Bioinformatics* 10:207.

Donald JE, Shakhnovich EI. 2009. SDR: a database of predicted specificity-determining residues in proteins. *Nucleic Acids Res.* 37:D191–D194.

Edwards RJ, Shields DC. 2004. GASP: gapped ancestral sequence prediction for proteins. *BMC Bioinformatics* 5:123.

Eisen JA, Fraser CM. 2003. Phylogenomics: intersection of evolution and genomics. *Science* 300:1706–1707.

Gao X, Vander Velden KA, Voytas DF, Gu X. 2005. SplitTester: software to identify domains responsible for functional divergence in protein family. *BMC Bioinformatics* 6:137.

Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem Sci.* 27:315–321.

Gu J, Wang Y, Gu X. 2002a. Evolutionary analysis for functional divergence of Jak protein kinase domains and tissue-specific genes. *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. 2001a. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol.* 18:453–464.
Gu X. 2001b. A site-specific measure for rate difference after gene duplication or speciation. *Mol Biol Evol.* 18:2327–2330.
Gu X. 2003. Functional divergence in protein (family) sequence evolution. *Genetica* 118:133–141.
Gu X. 2006. A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. *Mol Biol Evol.* 23:1937–1945.
Gu X. 2011. Statistical theory and methods for evolutionary genomics. Oxford: Oxford University Press.
Gu X, Li WH. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J Mol Evol.* 40:464–473.
Gu X, Vander Velden K. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* 18:500–501.
Gu X, Wang Y, Gu J. 2002b. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. *Nat Genet.* 31:205–209.
Gu X, Zhang J. 1997. A simple method for estimating the parameter of substitution rate variation among sites. *Mol Biol Evol.* 14:1106–1113.
Hannenhalli SS, Russell RB. 2000. Analysis and prediction of functional sub-types from protein sequence alignments. *J Mol Biol.* 303:61–76.
Huang Y, Zheng Y, Su Z, Gu X. 2009. Differences in duplication age distributions between human GPCRs and their downstream genes from a network prospective. *BMC Genomics* 10(1 suppl):S14.
Huang YF, Golding GB. 2011. Inferring sequence regions under functional divergence in duplicate genes. *Bioinformatics* 28:176–183.
Kalinina OV, Mironov AA, Gelfand MS, Rakhmaninova AB. 2004. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.* 13:443–456.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.
Knudsen B, Miyamoto MM. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc Natl Acad Sci U S A.* 98:14512–14517.
Landgraf R, Xenarios I, Eisenberg D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J Mol Biol.* 307:1487–1502.
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 G, Valencia A, Tress M. 2007. FireDB—a database of functionally important residues from proteins of known structure. *Nucleic Acids Res.* 35:D219–D223.
Lopez P, Casane D, Philippe H. 2002. Heterotachy, an important process of protein evolution. *Mol Biol Evol.* 19:1–7.
Mertz B, Gu X, Reilly PJ. 2009. Analysis of functional divergence within two structurally related glycoside hydrolase families. *Biopolymers* 91: 478–495.
Ohno S. 1970. Evolution by gene duplication. Berlin (Germany): Springer-Verlag.
Pazos F, Rausell A, Valencia A. 2006. Phylogeny-independent detection of functional residues. *Bioinformatics* 22:1440–1448.
Pazos F, Sternberg MJ. 2004. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A.* 101:14754–14759.
Pupko T, Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. *Proc Biol Sci.* 269:1313–1316.
Rausell A, Juan D, Pazos F, Valencia A. 2010. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci U S A.* 107:1995–2000.
Reva B, Antipin Y, Sander C. 2007. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* 8: R232.
Su Z, Gu X. 2012. Revisit on the evolutionary relationship between alternative splicing and gene duplication. *Gene* 504:102–106.
Su Z, Huang W, Gu X. 2011. Comment on "Positive selection of tyrosine loss in metazoan evolution." *Science* 332:917; author reply 917.
Su Z, Wang J, Yu J, Huang X, Gu X. 2006. Evolution of alternative splicing after gene duplication. *Genome Res.* 16:182–189.
Susko E, Inagaki Y, Field C, Holder ME, Roger AJ. 2002. Testing for differences in rates-across-sites distributions in phylogenetic subtrees. *Mol Biol Evol.* 19:1514–1523.
Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science* 278:631–637.
Wallace IM, Higgins DG. 2007. Supervised multivariate analysis of sequence groups to identify specificity determining residues. *BMC Bioinformatics* 8:135.
Wang X, Huang Y, Lavrov DV, Gu X. 2009. Comparative study of human mitochondrial proteome reveals extensive protein subcellular relocalization after gene duplications. *BMC Evol Biol.* 9:275.
Wang Y, Gu X. 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. *Genetics* 158:1311–1320.
Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387:708–713.
Xia X. 2011. Comparative genomics. In: Lu HH-S, Schölkopf B, Zhao H, editors. Handbook of computational statistics: statistical bioinformatics. Berlin: Springer. p. 567–600.
Zheng Y, Xu D, Gu X. 2007. Functional divergence after gene duplication and sequence-structure relationship: a case study of G-protein alpha subunits. *J Exp Zool B Mol Dev Evol.* 308:85–96.
Zhou H, Gu J, Lamont SJ, Gu X. 2007. Evolutionary analysis for functional divergence of the toll-like receptor gene family and altered functional constraints. *J Mol Evol.* 65:119–123.
Zou Y, Huang W, Gu Z, Gu X. 2011. Predominant gain of promoter TATA box after gene duplication associated with stress responses. *Mol Biol Evol.* 28:2893–2904.
Zou Y, Su Z, Huang W, Gu X. 2012. Histone modification pattern evolution after yeast gene duplication. *BMC Evol Biol.* 12:111.