Genetica **118:** 133–141, 2003.

Functional divergence in protein (family) sequence evolution

Xun Gu
Department of Zoology and Genetics, LHB Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA 50011, USA (Phone: +1-515-294-8075; Fax: +1-515-294-8457; E-mail: xgu@iastate.edu)

Key words: functional divergence, gene duplication, sequence evolution

Abstract

As widely used today to infer 'function', the homology search is based on the neutral theory that sites of greatest functional significance are under the strongest selective constraints as well as lowest evolutionary rates, and *vice versa*. Therefore, site-specific rate changes (or altered selective constraints) are related to functional divergence during protein (family) evolution. In this paper, we review our recent work about this issue. We show a great deal of functional information can be obtained from the evolutionary perspective, which can in turn be used to facilitate high throughput functional assays. The emergence of evolutionary functional genomics is also indicated. The related software DIVERGE can be obtained form http://xgu1.zool.iastate.edu.

Introduction

The goal of evolutionary functional genomics is to study the function of genes using molecular evolutionary approach (e.g., Golding & Dean, 1998). To this end, we have to clarify what kind of 'function' that can be predicted from sequence analysis. Under the population genetics model of molecular evolution (Kimura, 1983), 'function' can be quantified as the selection intensity *S*, which can apply at any level (residue, gene, etc.). That is, the 'evolutionary function' is defined as a parameter of population genetics that contributes to the fitness of the organism, which is related to function (biochemical, structural, or phenotypic) of a gene in complex ways. It involves interactions between the molecules and different levels of biological organization, including molecular complexes, pathways and cells, and eventually comprising individuals, populations, and species. The potential of evolutionary functional genomics is to generate experimentally testable hypotheses from genomic sequences (e.g., Lichtarge, Bourne & Cohen, 1996; Gaucher, Miyamoto & Benner, 2001; Wang & Gu, 2001).

In particular, we are focused on issues related to functional divergence after gene duplication (or speciation) (Fitch & Markowitz, 1970; Ohno, 1970; Li, 1983; Clark, 1994; Hughes, 1994; Fryxell, 1996; Nei, Gu & Sitnikova, 1997; Force et al., 1999). Since large value selection intensity *S* at an amino acid residue means functional importance (low evolutionary rate), and *vice versa* (Kimura, 1983), site-specific change in evolutionary rate (or selection intensity *S*) can be naturally interpreted as 'change of functional importance' (Gu, 1999).

Thanks to various statistical methods, it has been convincingly demonstrated that the evolutionary rate (the selection intensity) at a particular amino acid residue may vary during evolution (e.g., Miyamoto & Fitch, 1995; Lockhart et al., 1998; Gu, 1999, 2001; Lopez, Forterre & Philippe, 1999; Dermitzakis & Clark, 2001; Gaucher, Miyamoto & Benner, 2001; Jordan, Bishop & Gonzalez, 2001; Knudsen & Miyamoto, 2001). It means that during sequence evolution an amino acid residue can be changed from very conserved to highly variable, and *vice versa*. Indeed, some bioinformatics tools have been developed based on this principle (e.g., Casari, Sander & Valencia, 1995; Livingstone & Barton, 1996; Lichtarge, Bourne & Cohen, 1996; Landgraf, Fischer & Eisenberg, 1999).

Wang and Gu (2001) and Gu, Wang and Gu (2002) have proposed an evolutionary scenario that some kinds of functional divergence after gene duplication, as called 'type I' by Gu (1999), can cause changes of

evolutionary rate at some amino acid residues. Therefore, detection of site-specific rate shifts can provide a list of 'predicted' amino acid residues that may be responsible for functional divergence between member genes of a gene family. Some other studies support this 'functional divergence' view (e.g., Gaucher, Miyamoto & Benner, 2001; Jordan, Bishop & Gonzalez, 2001). The emphasis of this paper is to study functional divergence (biochemical-structural) of gene from the site-specific change of functional importance in protein sequence evolution.

Methods and software

The statistical model

Functional divergence of a protein (family) can occur after major evolutionary events such as gene duplication or speciation. Some of them may result in altered selective constraints (different evolutionary rates) at certain amino acid residues, which is called type I functional divergence, regardless of the underlying evolutionary mechanisms (Gu, 1999, 2001).

Recently many models are proposed to study the pattern of site-specific rate change in protein sequence evolution (e.g., Miyamoto & Fitch, 1995; Lockhart et al., 1998; Gu, 1999, 2001; Lopez, Forterre & Philippe, 1999; Dermitzakis & Clark, 2001; Gaucher, Miyamoto & Benner, 2001; Knudsen & Miyamoto, 2001). A brief description for ours models (1999, 2001) is as follows. Consider two gene clusters generated by a gene duplication event. Briefly speaking, in one cluster an amino acid site is called an F_1-site (functional divergence-related) if its evolutionary rate differs from that in the ancestral gene. If it is not the case, that is, the evolutionary rate (or selective constraint) at this site did not change during the process of duplication preservation and diversity, it is called an F_0-site (i.e., functional divergence-unrelated). As a result, for two gene clusters, a site can be in either of two states: (1) S_0 or (2) S_1, where S_0 is defined as a site being state F_0 in both clusters and S_1 is defined as a site being F_1 in at least one cluster. The coefficient of (type I) functional divergence, θ_{AB}, between two gene clusters A and B is defined as the probability of a site being in the S_1 state. Thus, a null hypothesis of $\theta = 0$ means that the evolutionary rate is virtually the same between duplicate genes at each site. Note that many models for rate variation among sites are the special case of $\theta = 0$. A probabilistic model has been developed (Gu, 1999, 2001) to estimate θ by establishing the relationship between this type of functional divergence and the observed amino acid configurations.

Rejection of the null hypothesis (i.e., $\theta > 0$) provides statistical evidence for altered selective constraints of amino acid sites after gene duplication. Thus, we can use a site-specific profile to identify responsible amino acid sites. Specifically we define $Q_k = P(S_1|X)$ to be the posterior probability that site k is in state S_1 ($0 \leq Q_k \leq 1$) when the amino acid configuration of this site is given. Large Q_k indicates a high possibility that the functional constraint (or, the evolutionary rate) of a site is different between two clusters. See the flow chart in Figure 1 for the illustration.

Functional distance analysis

One limitation of the two-cluster analysis described above is that it cannot test whether one duplicate gene has more shifted evolutionary rates than the other one. This issue is addressed by a simple method described below that can be applied when more than two homologous gene clusters are available (Wang & Gu, 2001). First, the functional distance between any two clusters is defined as $d_F = -\ln(1 - \theta)$. Under the assumption of independence, Wang and Gu (2001) have shown that d_F is additive, that is, for two clusters A and B, $d_F(A, B) = b_F(A) + b_F(B)$, where $b_F(x)$ is the functional branch length of a given gene cluster x. Large b_F value for a gene cluster indicates the evolutionary conservation may be shifted at many sites.

The estimated coefficients of type I functional divergence (θ) for all the pairs of clusters can be used to create a matrix of d_F values. Given this matrix, a standard least squares method can be implemented based on the formula $d_F(A, B) = b_F(A) + b_F(B)$ to estimate b_F for each gene cluster. If $b_F \approx 0$, it indicates that the evolutionary rate of each site in this duplicate gene has remained nearly the same since the gene duplication event, indicative that the derived state is more similar to the ancestral state for this particular cluster.

The software PHYBA: DIVERGE

The software system DIVERGE (DetectIng Variability in Evolutionary Rates among GEnes), which is available at the internet http://xgu1.zool.iastate.edu), follows this two-step procedure of statistical testing and then posterior predictions. It requires the user to input a multiple alignment of amino acid sequences,

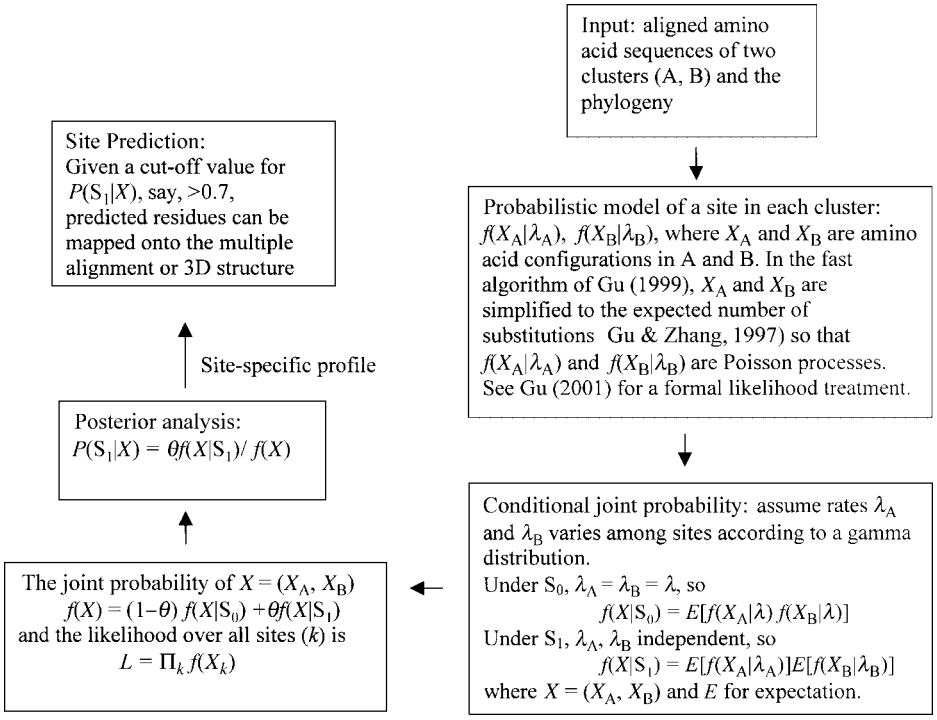


Figure 1. A flow chart to illustrate GU's methods (Gu, 1999, 2001).

in either FASTA or CLUSTAL format. Two options are available for the phylogenetic tree: (1) It can be input by the user using the PHYLIP format, standard to several software packages (e.g., PHYLIP, PAUP*, or CLUSTAL); (2) a neighbor-joining (NJ) tree (Saitou & Nei, 1987) can be generated and re-rooted by DIVERGE. Gene clusters of interests are selected by simply clicking the internal nodes of the tree. If multiple clusters are selected, DIVERGE performs the statistical analysis (Gu, 1999) for all pairs of clusters, as well as the site-specific profile (posterior analysis) to predict critical amino acid residues for functional divergence. If the user sets a cut-off value, usually >50%, residues with values above the cut-off will be highlighted in the multiple alignment. Then, the user may investigate the underlying mechanism for functional divergence by associating these sites with other biological evidence such as domain, motif, etc. In particular, DIVERGE provides a graphic tool to show the 3D protein structure, provided by a file in PDB format. Thus, predicted amino acid residues by the posterior probability analysis can be highlighted on the protein structure (if available). By allowing interactions with the protein structure, new discoveries about the interrelationships of residues, such as the spatial clustering of those with shifted rates, may be identified.

The performance of the above algorithm has been examined by several case studies (Gu, 1999; Naylor & Gerstein, 2000; Gaucher, Miyamoto & Benner, 2001; Jordan, Bishop & Gonzalez, 2001; Wang & Gu, 2001; Gu, Wang & Gu, 2002). For obtaining higher efficiency of detecting functional divergence-related residues, it is recommended to use a sequence dataset that satisfies the following conditions: (1) Each cluster must have at least four amino acid sequences. (2) Except for a large number of sequences, one should be cautious about the result when all pairwise sequence identities are >90%, because of the lacking of statistical power. And (3) multiple alignment should be reliable.

Applications for protein (family) evolution

Site-specific rate shifts is a general pattern after major evolutionary events

Our model is based on the principle that site-specific rate difference is the consequence of (type I) functional divergence of protein sequences. Here we show that the pattern site-specific rate shift (i.e., $\theta > 0$

Table 1. The pairwise coefficients of type I functional divergence (θ) of 13 mitochondria genes[a]

Gene	Sites[b]	Vertebrate versus invertebrate	Mammal versus bird	Mammal versus fish	Bird versus reptile
Atp6	203	0.21 ± 0.05	0.27 ± 0.07	0.34 ± 0.04	0
Atp8	46	0.30 ± 0.07	0.28 ± 0.19	0.42 ± 0.08	0
Co1	508	0.14 ± 0.03	0.31 ± 0.12	0.29 ± 0.05	0
Co2	224	0.06 ± 0.03	0.14 ± 0.06	0.26 ± 0.06	0.21 ± 0.15
Co3	256	0.18 ± 0.03	0.27 ± 0.07	0.20 ± 0.04	0
CytB	366	0.16 ± 0.02	0.22 ± 0.06	0.13 ± 0.02	0
Nd1	302	0.17 ± 0.03	0.14 ± 0.06	0.11 ± 0.03	0
Nd2	325	0.21 ± 0.03	0.29 ± 0.06	0.32 ± 0.03	0
Nd3	61	0.05 ± 0.07	0.51 ± 0.11	0.40 ± 0.07	0.35 ± 0.33
Nd4	438	0.23 ± 0.02	0.13 ± 0.06	0.21 ± 0.03	0
Nd4l	86	0.13 ± 0.09	0.32 ± 0.16	0.25 ± 0.08	0
Nd5	522	0.15 ± 0.02	0.24 ± 0.05	0.11 ± 0.02	0
Nd6	143	0.42 ± 0.06	0.60 ± 0.12	0.49 ± 0.06	0

[a] $\theta_{AB} \pm$ se: the coefficient of type I functional divergence between clusters A and B, and its standard error.
[b] Sites: the total number of amino acid residues in alignment.

significantly) is general during protein sequence evolution.

Vertebrate mitochondrial genomes show functional divergence after speciation. Mitochondrial genomes in vertebrates, composed of 13 genes, are good candidates for studying the impact of speciation events on protein functional divergence. Table 1 shows the estimated coefficients of functional divergence (θ) of 13 genes between different animal species. For most pairwise comparisons, θ values are significantly greater than 0, suggesting that altered functional constraints may take place at some amino acid residues after speciation. Interestingly, θ values between birds and reptiles are not significantly larger than 0, indicating a similar functional constraint between these two groups.

Vertebrate tissue-specific gene families show functional divergence after gene duplication. We have conducted a statistical analysis on 12 two-cluster vertebrate gene families. Table 2 shows that the coefficients of functional divergence (θ) between two gene clusters. As vary from 0.30 to 0.85, all of them are significantly greater than 0, implying that the pattern of altered selected constraints is common after gene duplication.

JAK proteins as an example for functional divergence after domain shuffling. Domain duplication (shuffling) is one of the major mechanisms for generating multiple-domain proteins. JAK gene family, a set of non-receptor tyrosine kinases, has two homologous kinase domains: a tandem kinase domain (functional) and a pseudokinase domain (function unknown). The inferred NJ tree of kinase domains suggests that JAK tandem kinase and pseudokinase domains are evolutionary distinct (Figure 2(A)). Interestingly, the coefficient of functional divergence (θ) between JAK tandem kinase and pseudokinase domains is 0.412 ± 0.049, suggesting that a significant altered functional constraint had occurred between them.

Important amino acid residues for functional divergence

To further explore the pattern of functional divergence after domain shuffling, the posterior analysis was conducted to define important amino acid residues responsible for altered functional constraints between these JAK tandem kinase and pseudokinase domains. Among 212 amino acid residues, 154 amino acid residues have almost no contribution to the functional divergence, that is, $P(S_1|X) < 0.5$. Figure 2(B) shows the 21 amino acid residues with the posterior probability to be functional divergence related greater than 0.9. These residues can be grouped into two categories: (1) conserved in the tandem kinase domain, whereas variable in the pseudokinase domain; (2) conserved in the pseudokinase domain, whereas variable in the tandem kinase. Interestingly, the existence of the second

Table 2. The coefficients of functional divergence (θ) between paralogous clusters of 12 gene families

Gene family	Member gene[a]		Sites[b]	$\theta_{AB} \pm se$[c]
	A	B		
Eph ligand family	EPLG7(5)	EPLG6(4)	190	0.30 ± 0.10
Y-box binding protein	YB-1a(13)	Variant A(5)	179	0.31 ± 0.16
Early growth response protein	EGR2(5)	EGR1(4)	310	0.36 ± 0.08
Stanniocalcin	STC(5)	STC2(4)	244	0.41 ± 0.19
CC chemokine receptor 2/5	CCR5(15)	CCR2(4)	341	0.42 ± 0.12
Sarcoplasmic reticulum Ca^{2+}-ATPase	SERCA1(13)	SERCA2(5)	990	0.49 ± 0.09
Endothelin	END1(6)	END2/3(6)	130	0.56 ± 0.17
Protein phosphatase 2C	PP2C-α(6)	PP2C-β(4)	323	0.57 ± 0.28
Succinyl-CoA synthetase β subunit	GTP-type(4)	ATP-type(4)	391	0.59 ± 0.14
Calponin	H1(11)	H2(5)	182	0.61 ± 0.15
Deleted in colorectal cancer	Neogenin(4)	DCC(4)	1335	0.74 ± 0.06
Lim-kinase	Limk2(6)	Limk1(4)	572	0.85 ± 0.08

[a] The number in parentheses is the number of sequences in each gene cluster.
[b] Sites: the total number of amino acid residues in alignment.
[c] $\theta_{AB} \pm$ se: the coefficient of type I functional divergence between clusters A and B, and its standard error.

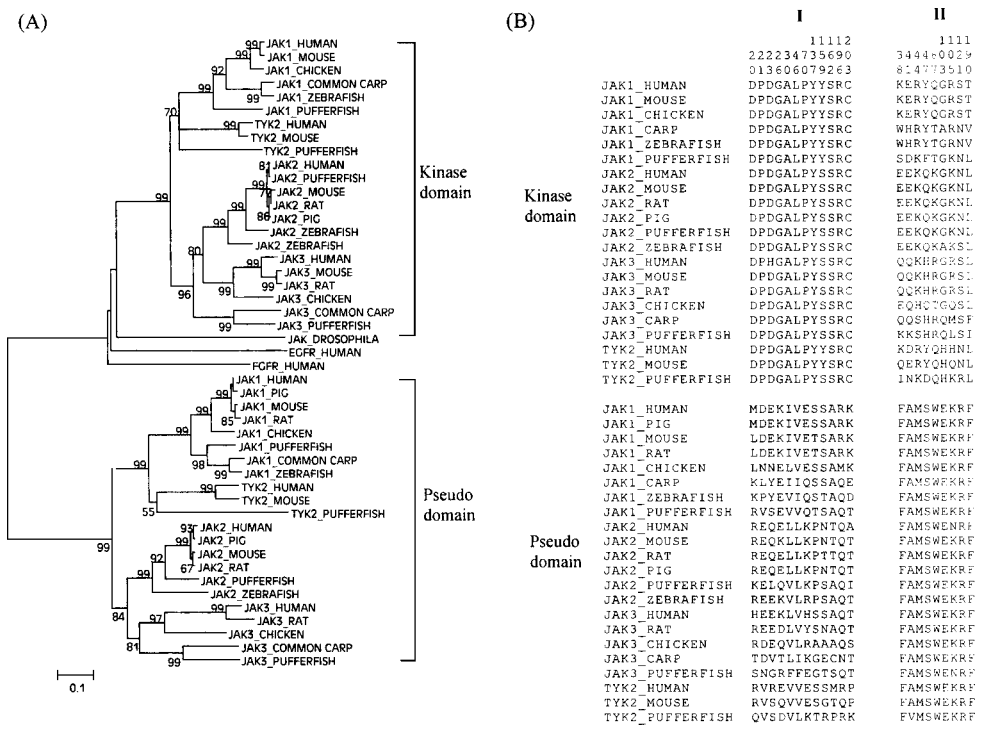


Figure 2. (A) The NJ tree of Jak tandem kinase domains and pseudokinase domains. (B) Functional divergence related amino acid residue candidates [$P(F_1|X) > 0.90$]: (I) conserved in tandem kinase domains, variable in pseudokinase domains; (II) conserved in pseudokinase domains, variable in tandem kinase domains.

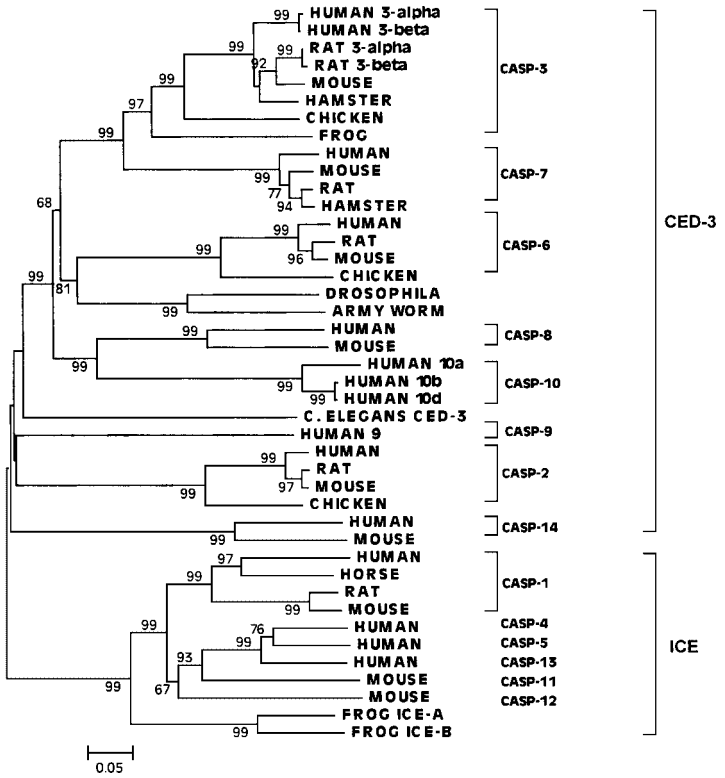


Figure 3. The phylogenetic tree of the caspase gene family, inferred by the NJ method based on the amino acid sequence with Poisson correction. Bootstrap values >50% are presented.

group implies that the pseudokinase domain in JAK may have some unique new functions (Gu, Wang & Gu, 2002).

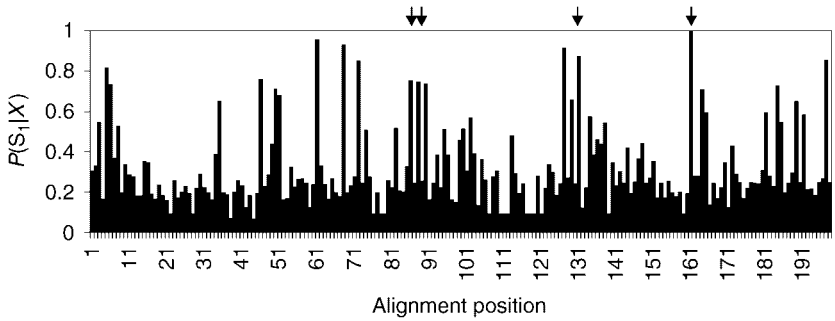
Structural basis for functional divergence in caspase families

The cascade of caspases (cysteine aspartyl proteases) is the key component in the apoptotic machinery, or programmed cell death. To date, at least 14 members of caspase gene family have been identified in mammals, which can be classified into CED-3 and ICE subfamilies. It has shown that the CED-3 type caspases are essential for most apoptotic pathways, whereas the major function of the ICE type caspases is to mediate immune response. X-ray crystallography has also shown a significant structural difference between these two types of caspases.

Based on the phylogeny of caspase family in Figure 3 inferred by the neighbor-joining (NJ) method, Wang and Gu (2001) have shown that the altered functional constraint between CED-3 and ICE subfamilies is statistically significant; the coefficient of functional divergence is $\theta = 0.29 \pm 0.05$. The site-specific profile for predicting critical amino acid residues is given by Figure 4(A). Among 198 residues, there are 20 amino acid residues corresponding to the cut-off value $P(S_1|X) > 0.67$, which are mapped onto the 3D structure of caspases. The resolved X-ray crystal structures of human caspases-1 and -3 have been used to illustrate the structural features of ICE and CED-3 subfamilies, respectively.

Wang and Gu (2001) have collected experimental evidence for four predicted residues that are involved in the functional–structural divergence between CED-3 and ICE subfamilies (Figure 4(B)). For example, amino acid residue 161 is critical for CED-3 caspase substrate specificity by interacting with a unique surface loop. At this position, all 22 sequences from the CED-3 subfamily contain an invariant tryptophan (W), whereas a variety of residues are present in the ICE subfamily. Crystal structural analysis reveals that W348 is a key determinant for the caspase-3 (CED-3) specificity because W348 forms a narrow pocket with the surface loop that is highly conserved in the CED-3 subfamily. Since the surface loop shared with CED-3

(A)



(B)

Site		CED-3	ICE
161	Sequence conservation	An invariant Trp (W)	Highly variable
	Structural features	Form a narrow pocket with an extra loop; form a H-bond	No extra loop; a shallow depression found
	Substrate specificity	Network with a group of amino acids; Hydrophilic side chains	Hydrophobic side chains
86/88	Structural features	No surface loop	Lie in an extra surface loop
131	Sequence conservation	Highly variable	Highly conserved
	Structural features	Not a cleavage site	Cleavage site for proenzyme processing

Figure 4. (A) The site-specific profile for predicting critical amino acid residues responsible for the functional divergence between CED-3 and the ICE subfamilies, measured by the posterior probability of being functional divergence-related at each site [$P(S_1|X)$]. The arrows point to four amino acid residues at which functional divergence between two subfamilies has been verified by experimentations. (B) Four predicted sites that have been verified by experimentation.

caspases is deleted in almost all ICE-type caspases, the high variability at this position of ICE subfamily is likely due to this 3D structure difference.

Isoform-specific functional divergence of Jak gene family

The NJ tree of four tissue-specific Jak gene members (Jak1, Jak2, Jak3 and Tyk2) in vertebrates indicates that Jak member genes were generated by two gene duplications in the early stage of vertebrates, that is, before the emergence of teleosts. The first gene duplication resulted in the common ancestor of Jak2/3 and Tyk2/Jak1, followed by the second one resulting in the current four member genes (Gu, Wang & Gu, 2002).

We have estimated θ between Jak1, Jak2 and Jak3 and found that the θ value varies between them, indicating overall no significant site-specific shift of evolutionary rate between them. To further explore the pattern of (site-specific) altered selective constraints during Jak family evolution, functional distance analysis is conducted for the whole sequence as well as three separate regions: (1) the tandem kinase domain (JH1), (2) the pseudokinase domain (JH2), and (3) the surrounding region excluding the JH1 and JH2 domains. As shown in Table 3, the level of altered selective constraints of member genes, measured by b_F, follows b_F(Jak1) > b_F(Jak2) > b_F(Jak3), while the level of altered selective constraints of domains follows JH1 > surrounding region > JH2. In particular, the b_F for Jak3 is virtually zero for all three domains, whereas JH2 (the pseudokinase domain) shows no significant functional branch length in Jak2 and Jak3.

Table 3 includes the ratios of nonsynonymous to synonymous rates (dn/ds) of Jak member genes, based on human–mouse orthologous members of a gene, which can be used to measure the difference of selective constraints among member genes of a gene family (Tsunoyama & Gojobori, 1998). Interestingly, Jak3,

Table 3. Functional branch lengths (b_F) of different regions in three Jak isoforms

	JH1	JH2	Others	Full length	dn/ds (full length)
Jak1	0.30	0.15	0.26	0.19	0.07
Jak2	0.28	0.10	0.09	0.04	0.09
Jak3	−0.03	−0.09	−0.07	−0.02	0.27

which has virtually $b_F = 0$, shows the highest dn/ds ratio. This observed negative association between dn/ds and b_F can be interpreted as follows. In the early stage after gene duplication, the functional divergence may occur in one of the two lineages (measured by large b_F value). If this process leads to the acquisition of some new functions, a stronger functional constraint (measured by low dn/ds value) is expected.

Discussion and conclusion

We have shown that site-specific rate shift (i.e., type I functional divergence) is a general pattern during the evolution of protein sequences. Moreover, the methodology Gu (1999, 2001) developed provides a useful evolutionary approach for understanding gene function in the era of genomics.

In addition to our methodology (Gu, 1999, 2001), there are several other methods available that are related to functional-divergence from molecular evolutionary analysis (e.g., Miyamoto & Fitch, 1995; Lockhart et al., 1998; Lopez, Forterre & Philippe, 1999; Dermitzakis & Clark, 2001; Gaucher, Miyamoto & Benner, 2001; Knudsen & Miyamoto, 2001). As each method has its own power and limitation, it actually regards to determine which method is suitable for a given dataset. On the other hand, there are many approaches to study function of proteins based on various evolutionary models (e.g., Lichtarge, Bourne & Cohen, 1996; Golding & Dean, 1998; Pollock, Taylor & Goldman, 1999; Suzuki & Gojobori, 1999). Since each only gets a particular aspect of the very complicated relationship between evolution, function and structure, an appropriate combination in practice is essential.

After gene duplication, there are two possibilities resulting in rate-difference (or selective constraint difference) of a site between duplicate genes. (1) One site becomes more conserved in one gene copy as a consequence of acquired new functions, or (2) it becomes more variable in one gene copy as a consequence of functional relaxation (e.g., via loss-of-function). It would be very interesting to distinguish between these two possibilities (Jordan, Bishop & Gonzalez, 2001), which will be the further direction of our study.

Acknowledgement

The study is support by the NIH grant RO1 GM62118.

References

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

Clark, A.G., 1994. Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. USA 91: 2950–2954.

Dermitzakis, E.T. & A.G. Clark, 2001. Differential selection after duplication in mammalian developmental genes. Mol. Biol. Evol. 18: 557–562.

Fitch, W.M. & E. Markowitz, 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4: 579–593.

Force, A., M. Lynch, F.B. Pickett, A. Amores, Y.L. Yan & J. Postlethwait, 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.

Fryxell, K.J., 1996. The coevolution of gene family trees. Trends Genet. 12: 364–369.

Gaucher, E.A., M.M. Miyamoto & S.A. Benner, 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. PNAS 98: 548–552.

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

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

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

Gu, X. & K. Vander Velden, 2002. DIVERGE: phylogeny-based analysis for functional–structural divergence of a protein family. Bioinformatics (in press).

Gu, J., Y. Wang & X. Gu, 2002. Evolutionary analysis for functional divergence of Jak protein kinase domains and tissue-specific genes. J. Mol. Evol. (in press).

Hughes, A.L., 1994. The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. B Biol. Sci. 256: 119–124.

Jordan, K., G.R. Bishop & D.S. Gonzalez, 2001. Sequence and structural aspects of functional diversification in class I-mannosidase evolution. Bioinformatics 17: 965–976.

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

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

Landgraf, R., D. Fischer & D. Eisenberg, 1999. Analysis of heregulin symmetry by weighted evolutionary tracing. Protein Eng. 12: 943–951.

Li, W.H., 1983. pp. 14–37 in Evolution of Genes and Proteins, M. Nei & R.K. Keohn. Sinauer Associates, Sunderland, MA.

Lichtarge, O., H.R. Bourne & F.E. Cohen, 1996. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257: 342–358.

Livingstone, C.D. & G.J. Barton, 1996. Identification of functional residues and secondary structure from protein multiple sequence alignment. Meth. Enzymol. 266: 497–512.

Lockhart, P.J., M.A. Steel, A.C. Barbrook, D.H. Huson, M.A. Charleston & C.J. Howe, 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. 15: 1183–1188.

Lopez, P., P. Forterre & H. Philippe, 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. 49: 496–508.

Miyamoto, M.M. & W.M. Fitch, 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. 12: 503–513.

Naylor, G.J. & M. Gerstein, 2000. Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. J. Mol. Evol. 51: 223–233.

Nei, M., X. Gu & T. Sitnikova, 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. PNAS 94: 7799–7806.

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

Pollock, D., W.R. Taylor & N. Goldman, 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J. Mol. Biol. 287: 187–198.

Rotonda, J., D.W. Nicholson, K.M. Fazil, M. Gallant, Y. Gareau, M. Labelle, E.P. Peterson, D.M. Rasper, R. Ruel, J.P. Vaillancourt, N.A. Thornberry & J.W. Becker, 1996. The three-dimensional structure of apopain/CPP32, a key mediator of apoptosis. Nat. Struct. Biol. 7: 619–625.

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

Suzuki, Y. & T. Gojobori, 1999. A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. 16: 1315–1328.

Tsunoyama, K. & T. Gojobori, 1998. Evolution of nicotinic acetylcholine receptor subunits. Mol. Biol. Evol. 15: 518–527.

Wang, Y. & X. Gu, 2000. Evolutionary patterns of gene families generated in the early stage of vertebrates. J. Mol. Evol. 51: 88–96.

Wang, Y. & X. Gu, 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics 158: 1311–1320.