Letter to the Editor

A Site-specific Measure for Rate Difference After Gene Duplication or Speciation

Xun Gu

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

Gene family proliferation by gene (genome) duplication has provided the raw materials for functional innovations (Ohno 1970; Lundin 1993; Holland et al. 1994; Henikoff et al. 1997; Golding and Dean 1998). Several models were proposed for functional divergence among member genes (e.g., Li 1983; Clark 1994; Hughes 1994; Fryxell 1996; Nei, Gu, and Sitnikova 1997; Force et al. 1999), but the details remain largely unknown. Gu (1999) developed a statistical method for testing type I functional divergence, i.e., changes in protein function between two gene clusters result in changes in selective constraints (and therefore shifted evolutionary rates) at some residues. It stands in contrast to type II functional divergence, i.e., changes in protein function between two gene clusters do not alter the level of selective constraints. Amino acid residues with rate shifts are the sites that have either gained or lost importance as a consequence of the change of function during evolution, as proposed by the hypothesis of type I functional divergence (Gu 1999). Moreover, type I functional divergence provides a biological basis for the covarion theory of molecular evolution (Fitch and Markowitz 1970).

If a statistical testing shows a significant rate difference between two gene clusters, it is of great interest to predict important amino acid residues, which can be further verified by available functional-structural evidence (Dermitzakis and Clark 2001; Gaucher, Miyamoto, and Benner 2001; Wang and Gu 2001). The posterior probability of each site is suitable to develop a statistically sound profile for selecting critical amino acid residues (Gu 1999), but little information is provided about how much rate difference is generated at these sites after gene duplication. In this article, we report a new site-specific profile for the rate difference, which is useful for studying the pattern of protein sequence evolution.

Consider two gene clusters generated by gene duplication (or speciation), e.g., see figure 1, the bone morphogenetic proteins (BMP) gene family tree. In each cluster, a site can be in either of two states: (1) F_0, which means no altered functional constraint after gene duplication, and (2) F_1, which means altered functional constraint at this site after gene duplication. As a result, there are four combined states in the case of two gene clusters: (1) F_0 in both clusters, denoted by $S_0 = (F_0, F_0)$, resulting in no rate difference between clusters; and (2) F_1 in at least one cluster, denoted by $S_1 = (F_0, F_1)$, (F_1, F_0), or (F_1, F_1), resulting in a rate difference between clusters. S_0 and S_1 are also called functional divergence configurations (Gu 2001). Let $P(S_1) = \theta$ and $P(S_0) = 1 - \theta$ be the probabilities of S_1 and S_0, respectively; θ is called the coefficient of (type I) functional divergence. Given these notations, the model of Gu (1999) model can be briefly described as follows.

- First, at a given site, the number of substitutions, X_1 (or X_2) $= i$, in clusters 1 (or 2), follows a Poisson distribution, denoted by $p(i|\lambda)$, whereas the evolutionary rate ($\lambda = \lambda_1$ or λ_2) varies among sites according to a gamma distribution $\phi(\lambda)$.
- Second, λ_1 and λ_2 are independent under S_1, whereas $\lambda_1 = \lambda_2$ under S_0. Let $X = (X_1, X_2)$. Thus, one can show that the joint (conditional) distributions of $X_1 = i$ and $X_2 = j$ is given by $P(X|S_1) = Q_1(i)Q_2(j)$, and $P(X|S_0) = K_{12}(i, j)$, respectively,

$$Q_1(i) = \frac{\Gamma(i + \alpha)}{i!\Gamma(\alpha)}\left(\frac{D_1}{D_1 + \alpha}\right)^i\left(\frac{\alpha}{D_1 + \alpha}\right)^\alpha$$

$$Q_2(j) = \frac{\Gamma(j + \alpha)}{j!\Gamma(\alpha)}\left(\frac{D_2}{D_2 + \alpha}\right)^j\left(\frac{\alpha}{D_2 + \alpha}\right)^\alpha$$

$$K_{12}(i, j) = \frac{\Gamma(i + j + \alpha)}{i!j!\Gamma(\alpha)}\left(\frac{D_1}{D_1 + D_2 + \alpha}\right)^i \times \left(\frac{D_2}{D_1 + D_2 + \alpha}\right)^j\left(\frac{\alpha}{D_1 + D_2 + \alpha}\right)^\alpha \quad (1)$$

where D_1 and D_2 are the total branch lengths in clusters 1 and 2, respectively, and α is the gamma distribution shape parameter (see eqs. 12 and 13 in Gu (1999) for details).

- Third, given the (prior) probability $P(S_1) = \theta$ and $P(S_0) = 1 - \theta$, the joint distribution of X_1 and X_2 can be expressed as:

$$P(X) = (1 - \theta)K_{12} + \theta Q_1 Q_2 \quad (2)$$

and a likelihood function can be built for estimating θ. When $\theta = 0$, equation (2) is reduced to a standard (homogeneous) model for rate variation among sites (e.g., Gu and Zhang (1997)).

- Fourth, estimation of θ requires a number of substitutions at each site for both gene clusters (i.e., X_1 and X_2). As X_1 and X_2 cannot be directly observed, a conventional solution is to use the number of minimum-required changes (m) as an approximation, which can be inferred by the parsimony under a known phylogenetic tree (Fitch 1971). However, m is biased because it does not consider the possibility of multiple hits. To solve this problem, Gu and Zhang (1997) developed an algorithm for estimating the expected number of substi-

Key words: protein sequence evolution, gene duplication, functional divergence, posterior rate difference.

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

Mol. Biol. Evol. 18(12):2327–2330. 2001

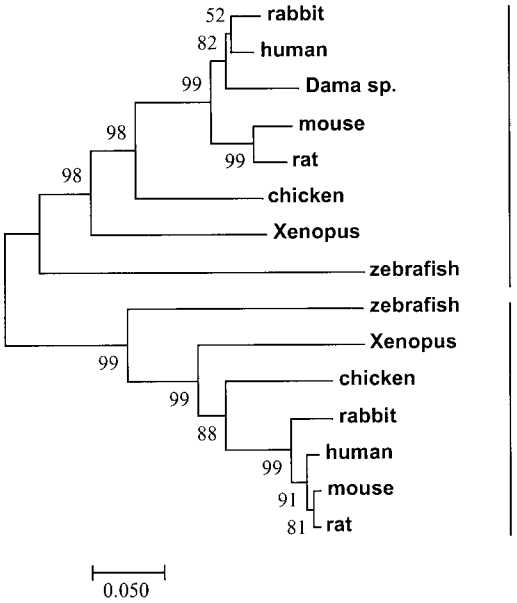

FIG. 1.—The phylogenetic tree of the BMP gene family, which was inferred by the neighbor-joining method, using amino acid sequences with Poisson distance. Bootstrapping values greater than 50% are presented.

tutions at each site, using a combination of ancestral sequence inference and maximum likelihood estimation.

● Fifth, the (site-specific) posterior probability for being S_1, i.e., type I functional-divergence related, is computed as follows:

$$P(S_1|X) = \frac{\theta Q_1 Q_2}{(1-\theta)K_{12} + \theta Q_1 Q_2} \quad (3)$$

Obviously, $P(S_1|X) = 0$ when $\theta = 0$, which is consistent with the standard model of rate variation among sites which assumes that the evolutionary rate of a site keeps constant during evolution, though it varies among sites.

Under the statistical framework described above, here we develop a site-specific measure for rate difference based on the posterior expectation. As $\lambda_1 = \lambda_2 = \lambda$ under S_0, the joint distribution of λ and $X = (X_1, X_2) = (i, j)$ is given by $P(\lambda, X|S_0) = p_1(i|\lambda)p_2(j|\lambda)\phi(\lambda)$, where $p_1(i|\lambda)$ and $p_2(j|\lambda)$ are the Poisson distributions of substitutions in clusters 1 and 2, respectively. Then, one can show that the conditional density of λ under S_0 is given by:

$$f(\lambda|X, S_0) = \frac{P(\lambda, X|S_0)}{P(X|S_0)} = \frac{p_1(i|\lambda)p_2(j|\lambda)\phi(\lambda)}{K_{12}(i, j)} \quad (4)$$

Under S_1, the evolutionary rates λ_1 and λ_2 are independent. Applying the Bayes theorem similar to the derivation of equation (4), one can show that the conditional density of λ_1 (or λ_2) under S_1 is given by, respectively,

$$f(\lambda_1|X, S_1) = \frac{P(\lambda_1, X_1|S_1)}{P(X|S_1)} = \frac{p_1(i|\lambda)\phi(\lambda_1)}{Q_1(i)}$$

$$\text{or } f(\lambda_2|X, S_1) = \frac{P(\lambda_2, X_1|S_1)}{P(X|S_1)} = \frac{p_2(j|\lambda)\phi(\lambda_2)}{Q_2(j)} \quad (5)$$

Then, by putting equations (1)–(5) together, we have obtained the posterior mean of rate under S_0 or S_1 as follows:

$$E[\lambda|X, S_0] = \int_0^\infty \lambda f(\lambda|X, S_0)\, d\lambda = \frac{i + j + \alpha}{D_1 + D_2 + \alpha}\bar{\lambda}$$

$$E[\lambda_1|X, S_1] = \int_0^\infty \lambda_1 f(\lambda_1|X, S_1)\, d\lambda_1 = \frac{i + \alpha_1}{D_1 + \alpha_1}\bar{\lambda}_1$$

$$E[\lambda_2|X, S_1] = \int_0^\infty \lambda_2 f(\lambda_2|X, S_1)\, d\lambda_2 = \frac{j + \alpha_2}{D_2 + \alpha_2}\bar{\lambda}_2 \quad (6)$$

where $\bar{\lambda}$, $\bar{\lambda}_1$, and $\bar{\lambda}_2$ are the mean rates of λ, λ_1, and λ_2 over sites, respectively.

Let v_1 (or v_2) be the evolutionary rate in cluster 1 (or 2). Under the current two-state model, we have $v_1 = \lambda$ under S_0, and $v_1 = \lambda_1$ under S_1. Thus, the expectation of v_1 can be expressed as $E[v_1] = P(S_0)\bar{\lambda} + P(S_1)\bar{\lambda}_1$. Therefore, the posterior mean of v_1 is given as follows:

$$E[v_1|X] = P(S_0|X)E[\lambda|X, S_0] + P(S_1|X)E[\lambda_1|X, S_1]$$
$$= P(S_0|X)\frac{i + j + \alpha}{D_1 + D_2 + \alpha}\bar{\lambda}$$
$$+ P(S_1|X)\frac{i + \alpha_1}{D_1 + \alpha_1}\bar{\lambda}_1 \quad (7)$$

For cluster 2, in the same manner we have:

$$E[v_2|X] = P(S_0|X)E[\lambda|X, S_0] + P(S_1|X)E[\lambda_2|X, S_1]$$
$$= P(S_0|X)\frac{i + j + \alpha}{D_1 + D_2 + \alpha}\bar{\lambda}$$
$$+ P(S_1|X)\frac{i + \alpha_2}{D_2 + \alpha_2}\bar{\lambda}_2 \quad (8)$$

Thus, a site-specific profile for rate difference can be defined as: $E[\Delta v|X] = E[v_1|X] - E[v_2|X]$, which is given by:

$$E[\Delta v|X] = P(S_1|X)\left(\frac{i + \alpha_1}{D_1 + \alpha_1}\bar{\lambda}_1 - \frac{j + \alpha_2}{D_2 + \alpha_2}\bar{\lambda}_2\right) \quad (9)$$

As these mean rates over all sites (i.e., $\bar{\lambda}_1$ and $\bar{\lambda}_2$) are usually unknown, we have to use the relative rate difference. For example, using $\bar{\lambda}_1$ as a reference, the relative rate is as follows:

$$r_k = P(S_1|X)\left(\frac{i + \alpha_1}{D_1 + \alpha_1} - \frac{j + \alpha_2}{D_2 + \alpha_2}c\right) \quad (10)$$

where $c = \bar{\lambda}_2/\bar{\lambda}_1$. In practice, c can be approximately estimated by the evolutionary distances using the same orthologous genes, i.e., the same evolutionary time.

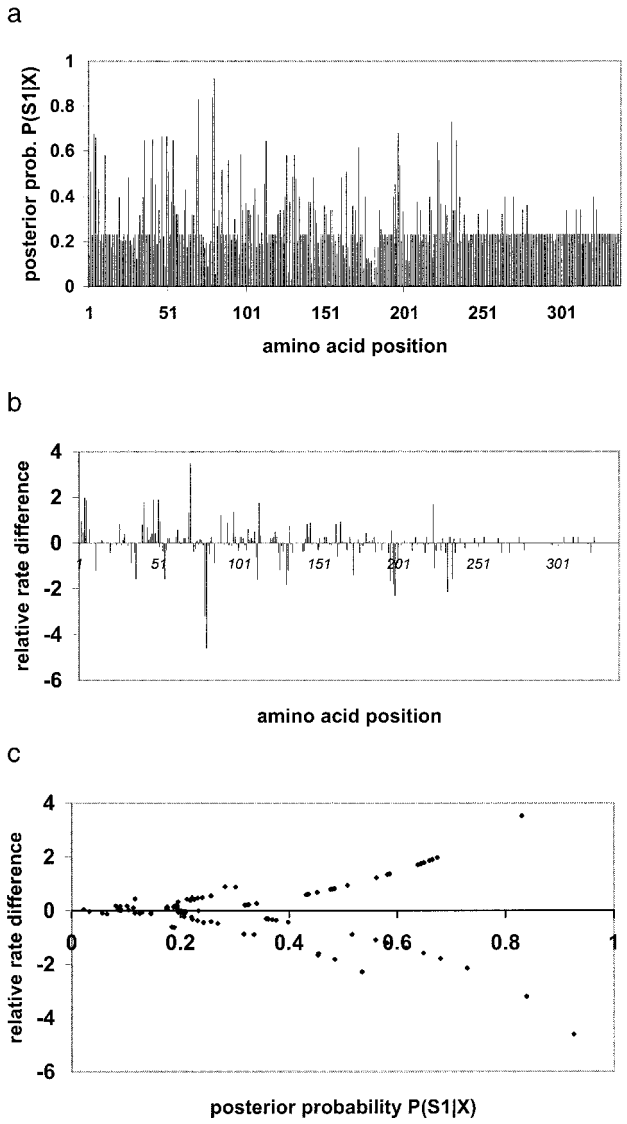

FIG. 2.—*a*, The site-specific profile for predicting critical amino acid sites that are responsible for the (type I) functional divergence, measured by the posterior probability. *b*, The site-specific profile for rate difference between *BMP2* and *BMP4*. Positive value means the rate of *BMP2* is higher than the rate of *BMP4*, and vice versa. *c*, The correlation between relative rate difference and posterior probability of (type I) functional divergence-related difference.

We use the vertebrate BMP gene family as an example. Figure 1 shows the phylogenetic tree of two BMP member genes, which is inferred by the neighbor-joining method (Saitou and Nei 1987). Apparently, *BMP2* and *BMP4* were generated by gene duplication in the early stage of vertebrates. Based on this topology, the coefficient of (type I) functional divergence is estimated to be $\theta = 0.283 \pm 0.067$, providing significant evidence for altered functional constraints between *BMP2* and *BMP4* after gene duplication.

The site-specific profile (posterior probability), $P(S_1|X)$, scores the relative likelihood of an amino acid residue to be involved in functional divergence between *BMP2* and *BMP4* (fig. 2, panel *a*). Among a total of 338 aligned sites, only a few sites receive high scores; for most of them, the score is only around 0.2. In particular, four sites have scores more than 0.7, whereas 12 sites have scores between 0.6 and 0.7. To understand the rate difference between *BMP2* and *BMP4* at these sites, we computed the site-specific profile of relative rate difference r_k (see fig. 2*b*); positive value means that the rate of *BMP2* is larger than that of *BMP4*, and vice versa if it is negative. It is expected that a site with large rate difference (positive or negative) should imply a high posterior probability given by equation (3). Indeed, figure 2*c* shows a strong correlation between these two measures.

Gu (2001) has developed a maximum likelihood framework for functional divergence, based on the Markov chain model. Using a similar approach, we can develop a site-specific profile for the rate difference. In this case, the posterior mean of rate difference (see eq. 10) should be expressed as follows:

$$E[\Delta v|X] = P(S_1|X)(E[\lambda_1|X, S_1] - E[\lambda_2|X, S_1]) \quad (11)$$

where $X = (X_1, X_2)$ is for the observed amino acid configuration at a site. $E[\lambda_1|X, S_1]$ (as well as $E[\lambda_2|X, S_1]$) can be computed under the framework of the Markov chain model (Yang 1997). The problem in practice is the computational time. Fortunately, our preliminary result shows that the performance of equations (10) and (11) is similar (unpublished data).

The methodology we developed (Gu 1999, 2001) provides a new approach for testing the site-specific rate difference after gene duplication or speciation. The current study provides a new site-specific profile for quantitatively measuring how much the functional constraint at a site can be changed after these evolutionary events, e.g., $P(S_1|X) = 0.93$ at site 80, indicating a strong rate shift pattern (type I functional divergence) between *BMP2* and *BMP4*. The relative rate difference at this site (-4.6) indicates a much stronger selective constraint in the *BMP2* gene than in the *BMP4* gene. Moreover, given the average evolutionary rate of $\sim 0.4 \times 10^{-9}$/year (using the human-mouse orthology with split time 100 MYA), the absolute (posterior) rate difference at this site can be computed as $-4.6 \times 0.4 \times 10^{-9} = -1.84 \times 10^{-9}$. Indeed, the evolutionary rate at this site is $\sim 2.03 \times 10^{-10}$ in *BMP2*, but $\sim 2.12 \times 10^{-9}$ in *BMP4*, indicating a ca. 10-fold rate change at site 80 after gene duplication. However, the rate in *BMP4* is not higher than the synonymous rate ($\sim 3 \times 10^{-9}$, estimated by human-mouse orthology). Though it is rough and indirect, the analysis indeed indicates that positive selection may not play an important role at this site.

In summary, this measure can provide a site-by-site basis for studying the relationship between the altered functional constraint and functional-structural assays, e.g., the effect of site-mutagenesis or the contribution of 3D difference. The functional-structural basis for type I functional divergence has been illustrated by Wang and Gu (2001). After gene duplication, there are two possibilities resulting in rate difference of a site between duplicate genes: (1) it 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). The sign (positive or negative) of the site-specific profile that indicates the trend of change in selective constraint is useful for understanding the underlying evolutionary mechanism.

In this report, we assume that the site-specific rate difference is equivalent to the site-specific altered functional constraint, which is valid as long as the mutation rate is not site specific. Different mutational rates owing to gene locations in the genome have virtually no effect on our analysis (Gu 1999).

Because biochemical properties (charge, hydrophobicity, etc.) of amino acid substitutions are not considered by this simple approach, the interpretation needs to be cautious in some cases, e.g., with many substitutions between amino acids R and K, which are both positively charged. This problem can be solved by two modifications. First, after a group of residues are selected, a follow-up checking based on some empirical rules may be informative. Second, we can improve our model to take this factor into account. For example, we can develop a weight matrix (or substitution matrix) of amino acid substitutions that is specific to each state (F_0 or F_1).

Using any given measure, it is not difficult to conduct a computational search to output a list of amino acid residues, each of which seems more conserved in one cluster than the other. However, we argue that statistical modeling and prediction is essential for several reasons. First, a simple list of these residues without statistical testing cannot be used as a piece of valid evidence to support or reject a scientific hypothesis, e.g., site-specific altered functional constraint after gene duplication. Second, the criterion for residue selection should be statistically sound. Third, for protein family sequences, a phylogeny-based profile is crucial to avoid any bias caused by unequal sequence sampling. For example, consider two gene clusters with an equal number of sequences. Cluster 1 includes closely related sequences, whereas cluster 2 includes distantly related sequences. Any prediction ignoring phylogeny can be misleading because many sites will show identical amino acid patterns in cluster 1. This problem (usually causing a high false positive rate) becomes serious for a large-scale analysis because visual inspection is not possible. At any rate, a statistically sound analysis is beneficial and cost-effective for functional and evolutionary genomics, as long as it is computationally fast.

Acknowledgments

This study was supported by NIH grant RO1 GM62118 to X.G. Accession numbers are as follows. BMP-2: AF041421, HUMBMP2A, DSPBMP2, MUSBMP2A, RNBMP2, GGBMP2, XLBMP2, D49971. BMP-4: MUSBMP4, RNBOMPR4A, HUMBMP4, AF042497, GGBMP4, XLBMP4, D49972.

LITERATURE CITED

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

DERMITZAKIS, E., and A. CLARK. 2001. Non-neutral diversification after duplication in mammalian developmental genes. Mol. Biol. Evol. **18**:557–562.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

FITCH, W. M., and 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, and 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, A., M. MIYAMOTO, and S. BENNER. 2001. Function–structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. Proc. Natl. Acad. Sci. USA **98**:548–552.

GOLDING, G. B., and A. M. DEAN. 1998. The structural basis of molecular adaptation. Mol. Biol. Evol. **15**:355–369.

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

———. 2001. Maximum likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. **18**: 453–464.

GU, X., and J. ZHANG. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol. Biol. Evol. **14**:1106–1113.

HENIKOFF, S., E. A. GREENE, S. PIETROKOVSKI, P. BORK, T. K. ATTWOOD, and L. HOOD. 1997. Gene families: the taxonomy of protein paralogs and chimeras. Science **278**(5338): 609–614.

HOLLAND, P. W. H., J. GARCIA-FERNANDEZ, N. A. WILLIAMS, and A. SIDOW. 1994. Gene duplication and the origins of vertebrate development. Development (**Suppl.**):125–133.

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

LI, W. H. 1983. Evolution of duplicated genes. *In* M. NEI and R. K. KOEHN, eds. Evolution of genes and proteins. Sinauer Associates, Sunderland, Mass.

LUNDIN, L. G. 1993. Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics **16**:1–19.

NEI, M., X. GU, and T. SITNIKOVA. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune systems. Proc. Natl. Acad. Sci. USA **94**: 7799–7806.

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

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

WANG, Y., and X. GU. 2001. Functional divergence in caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics **158**:1311–1320.

YANG, Z. 1997. PAML, a program package for phylogenetic analysis by maximum likelihood. CABIOS **13**:555–556.

NARUYA SAITOU, reviewing editor

Accepted August 2, 2001