**Personalized Monograph**

**2023v1**

# DIVERGE Analysis

**Xun Gu**

# Table of Contents

# Unit-1

# Type-I Functional Divergence after Gene Duplication: the Poisson-based Model

**Introduction**

Many organisms have undergone genome wide or local chromosome duplication events during their evolution (Ohno 1970). Consequently, many genes are represented as several paralogs in the genome with related but distinct functions. These gene family proliferations are thought to have provided the raw materials for functional innovations.

An understanding of the functional diversity of a gene family has been a major component in molecular evolutionary study. Extensive studies have been reported on the underlying mechanism of functional divergence after gene duplication. Ohno (1970) proposed that following gene duplication, one gene copy maintains the original function, while the other copy is free to accumulate amino acid changes as a result of functional redundancy or positive selection. Unless this type of functional divergence results in some new functions, over time all but one gene copy will be silenced by deleterious mutations. Hughes (1994) speculated that the ancestral gene might already be bifunctional and gene duplication simply allows each copy to specialize for one of several functions. Having realized the importance of coevolution between the interacted molecules (e.g., ligand/receptor), Fryxell (1996) argued that functional divergence may occur only when all genes in a pathway are duplicated simultaneously, e.g., by a genome duplication. It becomes clear that some evolutionary changes in the coding and/or regulatory regions after gene duplication must be responsible for the functional differences between members of a gene family.

An interesting question is whether we can identify these important amino acid (or nucleotide) sites; the methods for doing so may have great potential for functional genomics since they are cost-effective, and the predictions obtained can be further tested by experimentation. For example, one may infer amino acid sites that have experienced altered functional roles in a period of evolution. Gu (1999) developed a stochastic model for the functional divergence after gene duplication, which can estimate the level of functional divergence from sequence data, and predict important amino acid residues for these functional differences between member genes of a gene family. The method distinguish between these changes related to functional divergence and the background changes which mainly represent neutral evolution.

**Functional Divergence and Altered Functional Constraint**

A (homologous) gene cluster is defined as a monophyletic group of sequences under a phylogenetic tree. For example, two gene clusters are generated by an event of gene duplication, and each of them consists of several orthologous sequences (fig. 1*A*). It is commonly believed that after gene duplication, the evolutionary rate (l) at an amino acid site may increase and functional divergence may occur in the early stage, followed by the late stage, in which purifying selection plays a major role in maintaining related but distinct functions (fig. 1*B*). The underlying mechanism for this type of accelerated evolution after gene duplication is still in dispute. If the early-stage functional divergence occurred in one duplicate gene, changes of functional roles at the sites involved can be observed in the late stage. As a result, evolutionary rates at these sites are different between the two gene clusters. Such functional divergence, resulting in altered functional constraint, is called type I functional divergence.

The central tenet is that type I functional divergence after gene duplication is highly correlated with the change in evolutionary rate, which is analogous to a fundamental rule in molecular evolution: functional importance is highly correlated with evolutionary conservation (Kimura 1983). Alternatively, type II functional divergence does not result in different functional constraints between the two gene clusters, but evolutionary rates can be different between early and late stages (fig. 1B). For example, cluster-specific residues may be subject to this type of functional divergence. In this paper, we deal mainly with type I functional divergence; type II functional divergence will be discussed elsewhere. The relationship between functional divergence, altered functional constraint, and

evolutionary rate provides a theoretical basis for modeling the type I functional divergence during sequence evolution

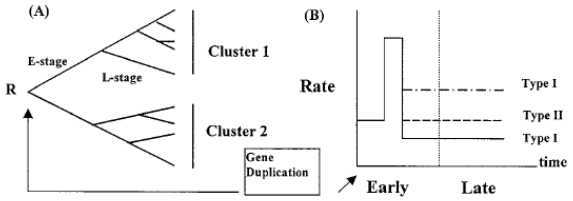



FIG. 1.—*A*, Two gene clusters after gene duplication; E and L represent early and late stages of gene cluster 1, respectively. *B*, Type I and type II functional divergences after gene duplication. In the early stage, the evolutionary rate (say, in cluster 1) may increase for functional-divergence-related change, but in the late stage, it may be higher (or lower) than the original rate, resulting in altered functional constraints between clusters 1 and 2 (type I functional divergence). If the rate in the late stage is the same as the original one again, no altered functional constraints between clusters 1 and 2 can be observed (type II functional divergence).

**A Simple ''Model-Free'' Method Rate Correlation between Two Gene Clusters**

If all sites have experienced no functional divergence after gene duplication, the two duplicate genes have no altered functional constraints, so the evolutionary rate of a site is always the same (or proportional) between them, i.e., the coefficient of rate correlation (over sites) is 1. Obviously, altered functional constraints caused by functional divergence will reduce the rate correlation. Consider a multiple alignment of amino acid sequences containing two gene family members (fig. 1). If orthologous sequences are functionally equivalent, the evolutionary rate (l) of a site remains constant (or proportional) among branches within a gene cluster, although it may vary among sites. Since a molecular clock is not assumed, lineage-specific factors such as generation time effect (Wu and Li 1985) will not affect the results. Hence, without loss of generality, the evolutionary rates in gene cluster 1 and gene cluster 2 are simply denoted by $\lambda_1$ and $\lambda_2$, respectively. The altered functional constraints between two gene clusters can be measured by the coefficient of rate correlation between $\lambda_1$ and $\lambda_2$,

$$r_\lambda = \frac{Cov(\lambda_1, \lambda_2)}{\sqrt{Var(\lambda_1)Var(\lambda_2)}} \quad (1)$$

where $Var(\lambda_1)$, $Var(\lambda_2)$ and $Cov(\lambda_1, \lambda_2)$ are the variances and covariance of $\lambda_1$ and $\lambda_2$, respectively. If there is no functional divergence after gene duplication, $\lambda_1=1$; otherwise, $\lambda_1<1$. Therefore, a convenient measure for functional divergence can be simply defined as

$$\theta_\lambda = 1 - r_\lambda \quad (2)$$

As $\theta_\lambda$ increases from 0 to 1, the functional divergence increases from very weak to extremely strong. In this sense, $\theta_\lambda$ is called the coefficient of functional divergence.

**The Poisson Model for Amino Acid Substitutions**

To avoid confusion, the term ''model-free'' means that there is no specific model for rate variation among sites and rate correlation between gene clusters; the method does require a model for amino acid changes at a site. A simple model is the Poisson process: at a given site, the number of amino acid changes ($X_i$, $i$=1, 2 for gene clusters 1 and 2, respectively) follows a Poisson distribution, i.e., the probability of $X_i=k$ is given by

$$p_i(k) = \frac{(\lambda_i T_i)^k}{k!} e^{-\lambda_i T_i}, \quad i = 1, 2 \quad (3)$$

where $T_1$ and $T_2$ are the total evolutionary times of clusters 1 and 2, respectively. In section A.1 of the appendix, we show that the coefficient of functional divergence defined by Eq.(2) is given by

$$\theta_\lambda = 1 - \frac{\sigma_{12}}{\sqrt{(V_1 - D_1)(V_2 - D_2)}} \quad (4)$$

where $D_1$ and $V_1$ (or $D_2$ and $V_2$) are the mean and variance of the number of changes (over sites) in cluster 1 (or cluster 2), respectively, and $\sigma_{12}$ is the covariance (over sites) between them.

To estimate $\theta_\lambda$ from equation (4), we need to know the number of changes at each site for each gene cluster (i.e., $X_1$ and $X_2$). Since $X_1$ and $X_2$ cannot be directly observed from the sequence data, a conventional solution is to use the minimum number of required changes ($m$) as an approximation, which can be inferred by the parsimony under a known phylogenetic tree (Fitch 1971). However, $m$ is a biased ''estimate'' for the true number of changes because it does not consider the possibility of multiple hits. This problem has been solved by using a combination of

ancestral sequence inference and maximum-likelihood estimation (Gu and Zhang 1997). Given a phylogeny, Gu and Zhang (1997) have shown that the expected number of changes ($X$) at a given site is the nonnegative solution of the likelihood equation

$$\sum_{i=1}^{M} \frac{\delta_i b_i}{1 - e^{-\hat{X} b_i / B}} = 1 \quad (5)$$

where $B$ is the total branch length of the gene cluster, and $b_i$ is the $i$-th branch length, $i$=1, . . . , $M$ ($M$ is the total number of branches); $\delta_i$= 1 if there is an amino acid change in the $i$-th branch, otherwise $\delta_i$=0. Extensive computer simulation has shown that the estimate of mean of expected number of changes, as well as that of variance, is asymptotically unbiased and robust against the accuracy of ancestral amino acid inference. Two interesting special cases are (1) $\hat{X} \approx m$ for short branch lengths, and (2) $\hat{X} = -M \ln(1-m/M)$ for equal branch lengths.

**Table 2**
**The Coefficients of Functional Divergence ($\theta_\lambda$) Between Gene Clusters by the Model-Free Method (eq. 4) and the Maximum-Likelihood Method Under the Two-State Model (MLE)**

| Genes | $N$ | Equation (4) | MLE |
|---|---|---|---|
| TF/LTF | 553 | 0.26 ± 0.08 | 0.19 ± 0.07 |
| TF/vTF | 553 | 0.13 ± 0.06 | 0.07 ± 0.03 |
| LTF/vTF | 553 | 0.07 ± 0.08 | 0.00 ± 0.03 |
| C-myc/N-myc | 276 | 0.52 ± 0.10 | 0.39 ± 0.08 |
| C-myc/L-myc | 276 | 0.57 ± 0.13 | 0.56 ± 0.12 |
| N-myc/L-myc | 276 | 0.39 ± 0.12 | 0.40 ± 0.12 |

NOTE.—$N$ is the total number of amino acid sites. See figures 2 and 6 for details on each gene cluster.

*Statistical Testing*

When the numbers of changes at each site in both clusters ($X_1$ and $X_2$) are obtained by Gu and Zhang's (1997) method, estimation of $\theta_\lambda$ is simple according to equation (4). Since $\theta_\lambda$>0 provides evidence for functional divergence after gene duplication, we have to test the statistical significance. Let $r_X$ be the coefficient of correlation between $X_1$ and $X_2$, which is defined by

$$r_X = \frac{\sigma_{12}}{\sqrt{V_1 V_2}} \quad (6)$$

Since $r_X$ reaches its maximum value $r_M$ when $\theta_\lambda$=0, i.e.,

$$r_X \le r_M = \sqrt{(1 - D_1/V_1)(1 - D_2/V_2)} \quad (7)$$

(see eq. A.9 in the appendix), the null hypothesis $H_0$: $\theta_\lambda$=0 is equivalent to $r_X = r_M$. As a standard coefficient of correlation, Fisher's transformation can be used to compute the confidence level of $r_X$:

$$z = 0.5 \ln\left(\frac{1+r}{1-r}\right)$$

Let $z_X$ and $z_M$, respectively, be the transforms of $r_X$ and $r_M$. The sampling variance of $z_X$ is approximately $V(z_X)$=1/($N$ -3), where $N$ is the sequence length. Under the null hypothesis ($r_X$=$r_M$), the $Z$ score [$Z$=($z_X$-$z_M$)/($N$-3)$^{0.5}$] approximately follows a normal distribution. For example, if the Z score is |$Z$| >1.96, the null hypothesis $\theta_\lambda$=0 can be rejected at the 5% significance level. Besides, by the delta method, the approximate sampling variance of can be computed as

$$Var(\hat{\theta}_\lambda) \approx \frac{1}{N-3}\left(\frac{1 - r_X^2}{r_M}\right)^2 \quad (8)$$

We should note that although $r_X$ is negatively correlated with $\theta_\lambda$ and useful for constructing a statistical test, it is not a good measure of the level of functional divergence because it is evolutionarily time-dependent (see eq. A.14 in the appendix)

*Example*s

Transferrins are iron-binding transport proteins which can bind two atoms of ferric iron $Fe^{3+}$. They are responsible for the transport of iron from sites of absorption and heme degradation to those of storage and utilization. There is only one gene in non-mammalian vertebrates (vTF). In mammals, two close-linked tissue-specific genes are found, which encode serum transferrin (TF) and lactotransferrin (LTF), respectively. Apparently, this gene duplication occurred before the radiation of mammals but after the divergence between birds and mammals. The results are summarized in Table 1.

**Table 1**
**Analysis of Functional Divergence Between the TF and LTF Gene Families Based on Equation (4), the Model-Free Estimate**

| Gene Clusters (1/2) | TF/LTF | TF/vTF | LTF/vTF |
|---|---|---|---|
| $D_1$ | 1.17 | 1.17 | 0.86 |
| $D_2$ | 0.86 | 2.20 | 2.20 |
| $V_1$ | 2.87 | 2.87 | 1.49 |
| $V_2$ | 1.49 | 4.24 | 4.24 |
| $\sigma_{12}$ | 0.76 | 1.59 | 1.04 |
| $r_X$ | 0.37 | 0.46 | 0.42 |
| $r_M$ | 0.50 | 0.53 | 0.45 |
| $\theta_\lambda$ | 0.26 ± 0.08 | 0.13 ± 0.06 | 0.07 ± 0.08 |
| $P$ | $<10^{-3}$ | <0.05 | >0.10 |

NOTE.—In the first case, TF represents cluster 1 and LTF represents cluster 2; in the second case, TF represents cluster 1 and vTF represents cluster 2; and in the third case, LTF represents cluster 1 and vTF represents cluster 2. See figure 2 for the definitions of these three clusters. $D_1$ and $V_1$ ($D_2$ and $V_2$) are the mean and variance of the number of changes in cluster 1 (cluster 2), respectively. $\sigma_{12}$ is the covariance and $r_X$ is the coefficient of correlation for the numbers of changes between gene clusters 1 and 2. $r_M$ is the expected value of $r_X$ when the evolutionary rate is completely correlated (i.e., $r_\lambda = 1$). The coefficient of rate correlation $\theta_\lambda$ is estimated according to equation (4), and the standard error is given by equation (8). The significance level ($P$ value) is computed by the method of Fisher's transformation.

**Table 2**
**The Coefficients of Functional Divergence ($\theta_\lambda$) Between Gene Clusters by the Model-Free Method (eq. 4) and the Maximum-Likelihood Method Under the Two-State Model (MLE)**

| Genes | $N$ | Equation (4) | MLE |
|---|---|---|---|
| TF/LTF | 553 | 0.26 ± 0.08 | 0.19 ± 0.07 |
| TF/vTF | 553 | 0.13 ± 0.06 | 0.07 ± 0.03 |
| LTF/vTF | 553 | 0.07 ± 0.08 | 0.00 ± 0.03 |
| C-myc/N-myc | 276 | 0.52 ± 0.10 | 0.39 ± 0.08 |
| C-myc/L-myc | 276 | 0.57 ± 0.13 | 0.56 ± 0.12 |
| N-myc/L-myc | 276 | 0.39 ± 0.12 | 0.40 ± 0.12 |

NOTE.—$N$ is the total number of amino acid sites. See figures 2 and 6 for details on each gene cluster.

## Two-state model for functional divergence

*The probabilistic model*

Consider an ideal case in which we already know exactly which sites are related to functional-divergence. Hence, all sites can be classified into either of two categories, $F_0$ (functional divergence-unrelated) and $F_1$ (functional divergence-related). In the $F_0$ category, the evolutionary rate (λ) of a site is the same between gene clusters, indicating no change in functional constraints. In contrast, the evolutionary rate of an $F_1$ site may have no correlation between gene clusters, because such sites have experienced altered functional constraints. However, in practice we do not know to which category each site belongs. This problem is solved by implementing a (two-state) probabilistic model: a given site can be in state $F_1$ with a probability of $P(F_1)$, or state $F_0$ with a probability of $P(F_0)$. Using the same notations as Eq.(1), we have $\mathrm{Cov}(\lambda_1, \lambda_2) = P(F_1)[\mathrm{Var}(\lambda_1)\mathrm{Var}(\lambda_2)]^{0.5}$ because $\mathrm{Cov}(\lambda_1, \lambda_2|F_0) = [\mathrm{Var}(\lambda_1)\mathrm{Var}(\lambda_2)]^{0.5}$ (completely correlated), and $\mathrm{Cov}(\lambda_1, \lambda_2|F_1)=0$ (independent). Then, one can show that

$$P(F_1) = 1 - r_\lambda = \theta_\lambda \qquad (9)$$

where $r_\lambda$ is the rate correlation between two gene clusters as defined by Eq.(1). That is, the coefficient of functional divergence ($\theta_\lambda$) can be interpreted as the probability of a site being in the state of functional divergence ($F_1$).

Denoting the probability of functional divergence at site $k$ by $\delta_k$, we mention that the current two-state model assumes that $\delta_k=1$ if it is $F_1$, otherwise $\delta_k=0$. Therefore, the expected proportion of sites to be functional divergence-related is given by $P(F_1)\times 1 + P(F_0)\times 0 = \theta_\lambda$. Furthermore, we assume that the evolutionary rate varies among sites according to a gamma distribution, i.e.,

$$\phi(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \qquad (10)$$

where $\lambda=\lambda_1$ or $\lambda_2$, respectively (Uzzel and Corbin 1971). The shape parameter α describes the degree of rate variation among sites, whereas $\beta$ is only a scalar. Since 1/α is the square of the coefficient of variation of $\lambda$, the larger the αvalue is, the weaker the rate variation is, and $\alpha=\infty$ means a uniform rate among sites.

The joint distribution of the number of changes, $P(X_1, X_2)$, can be derived as follows. For any $F_1$-site, the evolutionary rate is statistically independent between two clusters, whereas it is completely correlated at an $F_0$-site. Thus, the probability of $X_1=i$ in cluster 1 and $X_2=j$ in cluster 2 under state $F_0$ or $F_1$ is given by

$$\begin{aligned} P(X_1 = i, X_2 = j|F_1) &= Q_1(i)Q_2(j) \\ P(X_1 = i, X_2 = j|F_0) &= K_{12}(i, j) \end{aligned} \qquad (11)$$

respectively, where

$$\begin{aligned} Q_1(i) &= P(X_1 = i|F_1) = \int_0^\infty p_1(i)\phi(\lambda_1)d\lambda_1 \\ Q_2(j) &= P(X_2 = j|F_1) = \int_0^\infty p_2(j)\phi(\lambda_2)d\lambda_2 \\ K_{12} &= \int_0^\infty p_1(i)p_2(j)\phi(\lambda)d\lambda \end{aligned}$$

It is known that $Q_1(i)$ and $Q_2(j)$ are negative binomial distributions, i.e.,

$$\begin{aligned} Q_1(i) &= \frac{\Gamma(i+\alpha)}{i!\Gamma(\alpha)}\left(\frac{D_1}{D_1+\alpha}\right)^i\left(\frac{\alpha}{D_1+\alpha}\right)^\alpha \\ Q_2(j) &= \frac{\Gamma(j+\alpha)}{j!\Gamma(\alpha)}\left(\frac{D_2}{D_2+\alpha}\right)^j\left(\frac{\alpha}{D_2+\alpha}\right)^\alpha \end{aligned} \quad (12)$$

After some mathematical simplifications, one can show that $K_{12}(i, j)$ is given by

$$K_{12}(i,j) = \frac{\Gamma(i+j+\alpha)}{i!j!\Gamma(\alpha)}\left(\frac{D_1}{D_1+D_2+\alpha}\right)^i\left(\frac{D_2}{D_1+D_2+\alpha}\right)^j\left(\frac{\alpha}{D_1+D_2+\alpha}\right)^\alpha \quad (13)$$

Then, the joint distribution is given by $P(X_1, X_2)=P(F_0)P(X_1, X_2|F_0)+P(F_1)P(X_1, X_2|F_1)$, which can be expressed as

$$P(X_1, X_2) = (1-\theta_\lambda)K_{12} + \theta_\lambda Q_1 Q_2 \quad (14)$$

One can verify that the joint distribution $P(X_1, X_2)$ has the following properties: (*i*) The marginal distribution is a negative binomial distribution, i.e.,

$$\begin{aligned} P(X_1 = i) &= \sum_j P(X_1 = i, X_2 = j) = Q_1(i) \\ P(X_2 = j) &= \sum_i P(X_1 = i, X_2 = j) = Q_2(j) \end{aligned} \quad (15)$$

and (*ii*) the covariance between $X_1$ and $X_2$ is given by

$$\sigma_{12} = (1-\theta_\lambda)\frac{D_1 D_2}{\alpha} \quad (16)$$

*When one gene cluster has a single sequence*

If one cluster (say, cluster 2) has only one single sequence, the joint distribution of $X_1$ and $X_2$ needs to be modified since $X_2$ has only two states, $X_2$=0 or 1, with probabilities $Pr(X_2=0)=\exp(-\lambda_2 T_2)$ and $Pr(X_2=1)=1-\exp(-\lambda_2 T_2)$, respectively. In this case, the joint distribution of $X_1$ and $X_2$ at an $F_0$ site is $P(X_1=i, X_2=0|F_0)=K_{12}(i,0)$, and $P(X_1=i, X_2=1|F_0)=Q_1(i)-K_{12}(i,0)$. Similarly, the joint distribution of $X_1$ and $X_2$ at an $F_1$ site is $P(X_1=i, X_2=0|F_0)=Q_1(i)Q_2(0)$, and $P(X_1=i, X_2=0|F_1)=Q_1(i)[1-Q_2(0)]$. Then, one can show the joint distribution of $X_1$ and $X_2$ as follows

$$\begin{aligned} P(X_1 = i, X_2 = 0) &= (1-\theta_\lambda)K_{12}(i,0) + \theta_\lambda Q_1(i)Q_1(0) \\ P(X_1 = i, X_2 = 1) &= (1-\theta_\lambda)[Q_1(i) - K_{12}(i,0)] + \theta_\lambda Q_1(i)[1-Q_2(0)] \end{aligned} \quad (17)$$

*Maximum Likelihood Estimation (MLE)*

Let $P_k(i, j)$ be the probability of $X_1=i$ and $X_2=j$ at site $k$, Thus, the likelihood function can be expressed as

$$L(\mathbf{x}|data) = \prod_k P_k(X_1 = i, X_2 = j) \quad (18)$$

The parameter set $\boldsymbol{x}$ has four parameters, $D_1$, $D_2$, α and $\theta_\lambda$, which can be numerically estimated by a standard maximum likelihood approach. Since each marginal distribution follows a negative binomial distribution, we can first use Gu and Zhang (1997)'s method for estimating the mean and gamma shape parameter for each gene cluster, i.e., $D_1$, $\alpha_1$, and $D_2$, $\alpha_2$. Then, the initial value for α can be simply computed by $\alpha_0=(\alpha_1\alpha_2)^{0.5}$, and the initial value for $\theta_\lambda$ by the 'model-free' estimate [Eq.(4)]. Using these initial values, the ML estimates of $\theta_\lambda$ and α, as well as approximate sampling variances, can be obtained numerically. A likelihood ratio test (LRT) is constructed for testing the null hypothesis $H_0$: $\theta_\lambda=0$ *v.s.* $H_0$: $\theta_\lambda>0$. For the likelihood ratio $LR=\max\{L(H_0|\text{data})\}/\max\{L(H_A|\text{data})\}$, it is known that $-2\ln(LR)$ asymptotically follows a $\chi^2_{[1]}$. Some examples for MLE are shown in table 2. Generally speaking, ML estimates are slightly smaller than those of Eq.(4).

## Predicting Critical Amino Acid Residues

Our results (see tables 1 and 2) have provided strong statistical evidence for the functional divergence after gene duplication (i.e., $\theta_\lambda>0$). Therefore, it is of great interest to (statistically) predict which sites are likely to be responsible for these (type I) functional differences. Indeed, these sites can be further tested by using molecular, biochemical or transgenic approaches. We shall develop a site-specific profile for this purpose, which can be achieved by an empirical Bayesian model.

Remember that in the two-state model, each site has two possible states, $F_0$ (functional constraint) and $F_1$ (functional divergence), with the (prior) probabilities $P(F_1)=\theta_\lambda$ and $P(F_0)=1-\theta_\lambda$, respectively. To provide a statistical basis for predicting which state is more likely at a given site, we need to compute the (posterior) probability of state $F_1$ at this site with $X_1$ (and $X_2$) changes in cluster 1 (and 2), $P(F_1|X_1, X_2)$. Obviously, $P(F_0|X_1, X_2)=1-P(F_1|X_1, X_2)$. According to the Bayesian law and Eqs.(11) and (14), we can show

$$P(F_1|X_1, X_2) = \frac{P(F_1)P(X_1, X_2|F_1)}{P(X_1, X_2)} = \frac{\theta_\lambda Q_1 Q_2}{(1-\theta_\lambda)K_{12} + \theta_\lambda Q_1 Q_2} \quad (19)$$

Then, given $X_1=i$ and $X_2=j$, the posterior (probability) ratio can be defined as follows

$$R_{ij} = \frac{P(F_1|X_1=i, X_2=j)}{P(F_0|X_1=i, X_2=j)} = \frac{\theta_\lambda}{1-\theta_\lambda}\frac{Q_1(i)Q_2(j)}{K_{12}(i,j)} \quad (20)$$

which turns out to be

$$R_{ij} = \frac{\theta_\lambda}{1-\theta_\lambda}\frac{\Gamma(i+\alpha)\Gamma(j+\alpha)}{\Gamma(i+j+\alpha)}\left(1+\frac{D_2}{D_1+\alpha}\right)^i\left(1+\frac{D_1}{D_2+\alpha}\right)^j\left(1-\frac{D_1D_2}{(D_1+\alpha)(D_2+\alpha)}\right)^\alpha \quad (21)$$

We may use either Eq.(19) or Eq.(21) to identify these amino acid sites that may be responsible for the functional divergence, given a cut-off value. In practice, the choice of a cut-off value is somewhat arbitrary, from $P(F_1|X_1, X_2)>0.5$ ($R_{ij}>1$) to $P(F_1|X_1, X_2)>0.95$ (or $R_{ij}>20$).

## Appendix: some technical comments

*Derivation of Eq.(4)*

First we consider the Poisson process at a given site, in which the first and second moments can be expressed as the following conditional expectations,

$$\begin{aligned} E[X_i|\lambda_i] &= \lambda_i T_i \\ E[X_i^2|\lambda_i] &= \lambda_i T_i + (\lambda_i T_i)^2 \end{aligned} \quad (A\text{-}1)$$

($i$=1, 2). If there is no gene conversion or recombination between the two homologous genes, amino acid substitutions at a site are independent between two monophyletic gene clusters and therefore,

$$E[X_1X_2|\lambda_1, \lambda_2] = E[X_1|\lambda_1] \times E[X_2|\lambda_2] \quad (A\text{-}2)$$

The evolutionary rates ($\lambda_1$ and $\lambda_2$) are not only correlated but also different among sites, which in principle can be described by a general joint distribution, $\Phi(\lambda_1, \lambda_2)$. To compute the mean and variance over all sites (for each cluster), let $\varphi(\lambda_1)$ and $\varphi(\lambda_2)$ be the marginal distributions of $\Phi(\lambda_1, \lambda_2)$, which describe the rate variation among sites. By definition, they are given by

$$\begin{aligned} \phi(\lambda_1) &= \int_0^\infty \Phi(\lambda_1\lambda_2)d\lambda_2 \\ \phi(\lambda_2) &= \int_0^\infty \Phi(\lambda_1, \lambda_2)d\lambda_1 \end{aligned}$$

respectively. According to the conditional probability theory, one can show that

$$E[X_i] = E[E[X_i|\lambda_i]] = \int_0^\infty \lambda_i T_i \phi(\lambda_i)d\lambda_i = E[\lambda_i]T_i, \quad (A\text{-}3)$$

$i$=1,2, where $E[\lambda_i]$ is the mean rate of $\lambda_i$. In the same manner, we have

$$E[X_i^2] = E[\lambda_i]T_i + E[\lambda_i^2]T_i^2, \quad i = 1, 2 \tag{A-4}$$

For simplicity, let $D_i$=$E[X_i]$ and $V_i$=E[$X_i^2$]-( $E[X_i]$)$^2$. From Eq.(A.4) the variance of $\lambda_i$, Var(=E[$\lambda_i^2$]-( E[$\lambda_i$])$^2$, is given by

$$Var(\lambda_i) = (V_i - D_i)/T_i^2, \quad i = 1, 2 \tag{A-5}$$

Now consider the covariance between $\lambda_1$ and $\lambda_2$. From Eqs.(A.1) and (A.2) we have

$$E[X_1 X_2] = T_1 T_2 \int_0^\infty \lambda_1 \lambda_2 \Phi(\lambda_1 \lambda_2) d\lambda_1 d\lambda_2 = T_1 T_2 E[\lambda_1 \lambda_2] \tag{A-6}$$

and therefore the covariance between $X_1$ and $X_2$, $\sigma_{12}$, is given by

$$\sigma_{12} \;\; = \;\; T_1 T_2 \; Cov(\lambda_1, \lambda_2) \tag{A-7}$$

Then, from Eqs.(A.5) and (A.7), one can easily show that the coefficient of rate correlation $r_\lambda$ defined by Eq.(1) is given by

$$r_\lambda = \frac{\sigma_{12}}{\sqrt{(V_1 - D_1)(V_2 - D_2)}} \tag{A-8}$$

which directly leads to Eq.(4). Since $r_\lambda \leq 1$, we have

$$\sigma_{12} \leq \sqrt{(V_1 - D_1)(V_2 - D_2)}$$

which means

$$r_X = \frac{\sigma_{12}}{\sqrt{V_1 V_2}} \leq r_M = \sqrt{(1 - D_1/V_1)(1 - D_2/V_2)} \tag{A-9}$$

*A short note on rate variation among sites*

The gamma distribution model for rate variation among sites assumes no altered functional constraints during evolution, i.e., $\theta_\lambda$=0. Here we use a simple case to show that the estimation of the shape parameter α may be biased if the assumption of $\theta_\lambda$=0 is violated. In the two-cluster case (figure 1A), let $X$=$X_1$+$X_2$ be the (total) number of changes at a site. One can show that $X$ follows a negative binomial distribution if $\theta_\lambda$=0, i.e., no altered functional constraints (e.g., Gu and Zhang 1997). Under this model, the variance of $X$ is given by

$$V = D + \frac{D^2}{\alpha} \tag{A-10}$$

where $D$ is the mean of $X$. In the same manner for each cluster we have $V_1$=$D_1$+$D_1^2/\alpha$, and $V_2$=$D_2$+$D_2^2/\alpha$. On the other hand, we mention $X$=$X_1$+$X_2$ so that $D$=$D_1$+$D_2$ and $V$=$V_1$+$V_2$+2 $\sigma_{12}$. From Eq.(A.8) we have

$$V = V_1 + V_2 + 2(1 - \theta_\lambda)\sqrt{(V_1 - D_1)(V_2 - D_2)} \tag{A-11}$$

Therefore, if one defines α* as α*=$D^2$/($V$-$D$), one can easily show

$$\alpha^* = \frac{\alpha}{1 - b\theta_\lambda} \geq \alpha \tag{A-12}$$

where $b$=2$D_1D_2$/($D_1$+$D_2$)$^2$; α*=α only when $\theta_\lambda$=0. If we use the method of moments to estimate α under the assumption of no altered functional constraints between these two gene clusters, we obtain α*=$D^2$/($V$-$D$). According to Eq.(A.12), for the sufficient large number of sites, the following relation holds

$$E[\hat{\alpha}] \approx \alpha^* \geq \alpha \tag{A-13}$$

# Unit-2

# Type-I Functional Divergence after Gene Duplication: Caspase Gene Family Analysis

### Introduction

After gene duplication, the classical model (Ohno 1970) suggests that one gene copy maintains the original function, while the other copy is free to accumulate amino acid changes toward functional divergence. Since then, many specific models have been proposed (*e.g.*, Li 1983; Clark 1994; Force *et al.* 1999). However, the details of functional divergence between duplicate genes remain largely unexplored. Gu (1999) developed a method to detect amino acid residues that contribute to functional divergence after gene duplication, which can be considered as candidates for further experimentation. Certainly, its effectiveness for functional genomics needs to be verified by using gene families with substantial biological/structural information.

Apoptosis, or programmed cell death, is an ordered process in which cells commit suicide when they are not needed or are potentially harmful. The key component in the apoptotic machinery is a cascade of *c*ysteine *asp*artyl prote*ase* s (caspases). All caspases, which are initially inactive proenzymes, share the same processing scheme to achieve mature forms after cleavage(s) at specific Asp sites. To date, at least 14 members of the caspase gene family have been identified in mammals, which can be further classified into two major subfamilies, CED-3 and ICE. Substantial evidence has shown that the CED-3-type caspases are essential for most apoptosis pathways. In contrast, the major function of the ICE-type caspases is to mediate immune response, although some members may play a role in cell death in some circumstances. X-ray crystallography has also shown a significant structural difference between these two types of caspases. In this study, we take advantage of experimental evidence of caspases to study the functional-structural basis of statistical predictions from Gu's (1999) method.

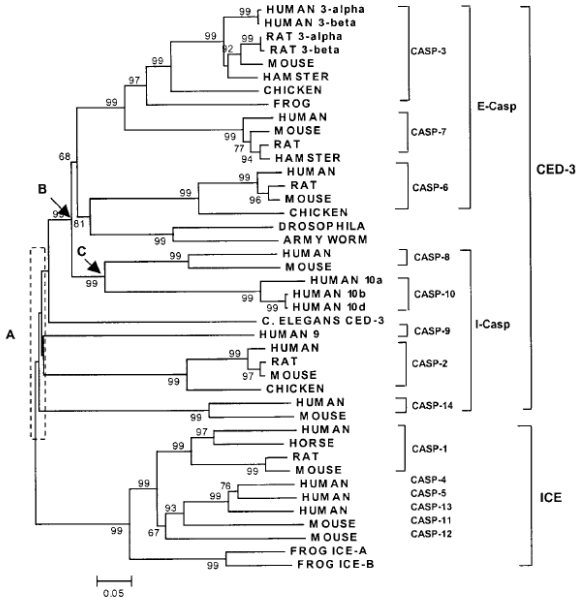



Figure 2.—The phylogenetic tree of the caspase gene family, inferred by the neighbor-joining method on the basis of the amino acid sequence

with Poisson correction. Bootstrap values _50% are presented. Initiator caspases (I-casps) are involved in upstream regulatory events, and effector caspases (E-casps) directly lead to cell disassembly.

### Evolution of caspase-mediated molecular pathways

The phylogenetic tree (Figure 2) of the caspase gene family was inferred by the neighbor-joining (NJ) method. The presence of caspases in vertebrates, arthropods, and nematodes suggests that the emergence of the caspase gene

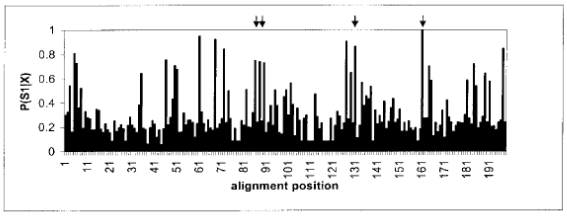
family might be close to or even earlier than the origin of the animal kingdom. Aravind et al. (1999) suggested that caspase may evolve from an ancient protease supergene family, but the root of the inferred tree (Figure 2) remains unclear. The evolutionary pattern of caspases can be generally described as follows. On the basin of the tree (see A in Figure 2), there were at least four duplication events that had occurred during a very short time period, resulting in five major lineages: (i) the ICE subfamily, consisting of caspase-1, -4, -5, -13, -11, and -12; (ii) caspase-14; (iii) caspase-2; (iv) caspase-9; and (v) the common ancestor of caspase-8/-10 and caspase-3/-6/-7. In addition, the effector caspases (E-casp-3/-7/-6) and the ancestor of caspase-8 and -10 were generated before the emergence of arthropods. Interestingly, in contrast to the major (ancient) lineages in CED-3-type caspases, ICE-type caspases diversified recently after the divergence of amphibians and mammals, and some of them (e.g., caspase-4 and -5) arose even after the mammalian radiation.

## Predicting critical residues for type I functional divergence between CED-3 and ICE

We estimated that the coefficient of functional divergence between ICE and CED-3 subfamilies is 0.29± 0.05 [the ML option in Gu's (1999) method], implying that the altered functional constraint between them is statistically significant. Further, we use the posterior probability P(S1|X) to predict critical amino acid residues responsible for type I functional divergence between CED-3 and ICE subfamilies. The baseline of the site-specific profile measured by P(S1|X) is 0.2–0.3 (Figure 4A). Thirty-two sites (16% of total sites) have P(S1|X)>0.5. The fact that most sites have scores 50% indicates their similar functional roles between CED-3 and ICE.

Although posterior analysis is widely used in bioinformatics, the cutoff value for residue selection is usually empirical. We found that when the first 21 highest-scored residues are removed from the multiple alignment, the estimate of $\theta$ is virtually 0. These 21 amino acid residues (among 198 residues) corresponding to the cutoff value P(S1|X) 0.61 are then chosen for further analysis. Of course, this procedure is meaningful only when $\theta$>0 significantly.

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 functionally 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 experimentation. (B) Four predicted sites that have been verified by experimentation.

## The functional-structural basis of altered functional constraints

We mapped these 21 predicted sites onto the 3-D structure of caspases. The resolved X-ray crystal structures of human caspase-1 and -3 (Wilson et al. 1994; Rotonda et al. 1996) were used to illustrate the structural features of ICE and CED-3 subfamilies, respectively. From the literature, we found experimental evidence for four predicted residues that are involved in the functional-structural divergence between CED-3 and ICE subfamilies (Figure 4B):

1. *Residue 161(348) (In the literature, this site is numbered as W348, according to the protein sequence of human caspase-1) is critical for CED-3 caspase substrate specificity by interacting with a unique surface loop in 3-D*

*structure [P(S1|X) =0.999.]* 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 (Figure 5). Crystal structural analysis reveals that W348 is a key determinant for the caspase-3 (CED-3)-type specificity. First, W348 forms a narrow pocket with the surface loop that is highly conserved in the CED-3 subfamily; see the boxed region in Figure 5. The steric constriction due to this pocket determines the preference of caspase-3 to the substrates with small hydrophilic side chains. Second, W348 along with a group of residues forms a hydrogen bond network, which affects the interaction with the substrate. In contrast, the surface loop shared with CED-3 caspases seems to be deleted in all ICE-type caspases, as shown in the boxed region in Figure 5. Hence, the relaxed evolutionary constraint observed at this position in the ICE subfamily is likely to be caused by the 3-D structural difference.

2. *Residues 86 [P(S1|X) =0.75] and 88[P(S1|X) =0.74] are responsible for 3-D difference with an unknown functional role.* Indeed, in human caspase-1 (ICE), these two residues appear to lie in a small loop that is not found in the CED-3 subfamily.

3. Residue 131 [P(S1|X)=0.866] is proteolytic site specific to the ICE subfamily. All caspases are synthesized as inactive proenzymes that need to be processed to the mature forms. However, distinct cleavage sites within the precursors are found for two subfamilies. D131 is known as a cleavage site in human caspase-1 (ICE type). All ICE-type caspases preserve an Asp (D) at this position, except for mouse caspase-12 (Asn, E). However, human caspase-3 (CED-3 type) utilizes two other Asn sites for cleavage (Rotonda et al. 1996) so that the functional role of position 131 in CED-3 caspases is no longer important. Therefore, the altered evolutionary constraints at this position can be well explained by the different utilization of cleavage sites for the precursor processing between CED-3 and ICE subfamilies.

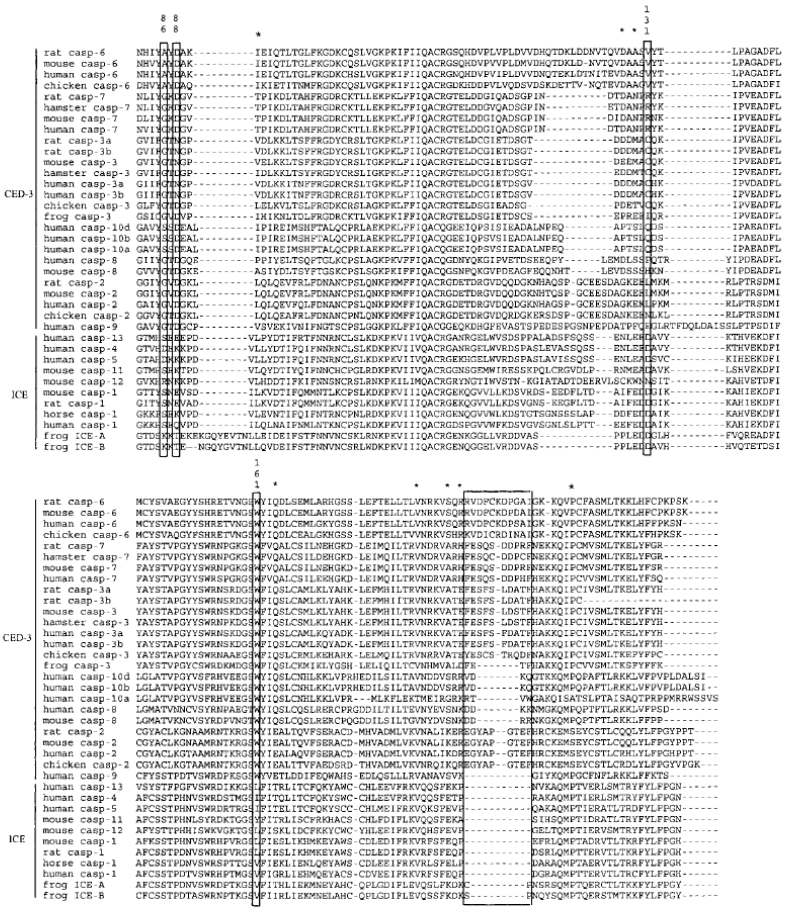



FIGURE 5.—Alignment of predicted regions of caspases. Four predicted sites with experimental evidence are highlighted. The sites with asterisks are predicted residues within this region. The boxed region in the C terminus is the critical region for CED-3 substrate specificity: Most CED-3-type caspases form a surface loop, whereas a shallow depression is found in ICE-type caspases.

## Pattern of type I functional divergence among CED-3-type caspases

The CED-3 subfamily consists of a specific group of caspases that mediate the programmed cell death in a well-regulated proteolytic cascade and employ related but distinct functions. Here we infer the trend of altered functional constraint of each cluster. Due to the data availability, we study five gene clusters: caspase-3, -7, -6, -8/-10, and -2. The upper diagonal of Table 1 shows pairwise coefficients of type I functional divergence ($\theta$) between them; all of them are significantly >0 ($P$<0.05), with only one exception; i.e., $\theta$=0 0.006 between caspase-7 and cluster -8/-10.

To explore the pattern of type I functional divergence in each cluster, we performed functional distance analysis (see methods). The pairwise functional distances ($d_F$) between clusters are shown in the lower diagonal of Table 1. The star-like tree presented in Figure 6 shows the type I functional branch length ($b_F$) of each cluster, estimated by the least-squares method. The null hypothesis of equal $b_F$ value for each cluster was statistically rejected ($P$<0.05).

Long functional branch lengths ($b_F$) of caspase-3, -6, and -2 suggest that these genes may have undergone extensive altered functional constraints as a result of specialized functional roles in apoptosis (Figure 6). Supportive experimental evidence is summarized as follows: (*i*) the non-redundant functional role of caspase-3 in neurological apoptosis is confirmed by caspase-3 -/- knockout mice; (*ii*) caspase-6 and -3 have different substrate specificity, but both participate in the protease amplification cycle by activating each other, which triggers a series of apoptotic interactions, and (*iii*) caspase-2 has its unique dual-role position in positive and negative regulation in apoptosis by differential expression of two alternative splicing isoforms. This dual-role property is also confirmed by knockout mice: Caspase-2 deficiency causes one defective apoptotic pathway (mediated by granzyme B and perforin) but accelerates another

pathway (cell death of motor neurons).

In contrast, virtually zero bF values of caspase-7 and -8/-10 indicate that the evolutionary rate of each site in these genes is almost identical to that of the ancestral gene. In this regard, these caspases may inherit a large component of ancestral function during caspase gene family evolution.

For each duplicate gene, the average intensity of functional constraints can be approximately measured by the $d_N/d_S$ ratio between the human and mouse. Interestingly, caspase-3, -6, and -2 (long $b_F$) have lower $d_N/d_S$ ratios than caspase-7 and -8/-10 (zero $b_F$), indicating that type I functional divergence in caspases may result in a stronger functional role in evolutionary novelties after gene duplication (Figure 6B).

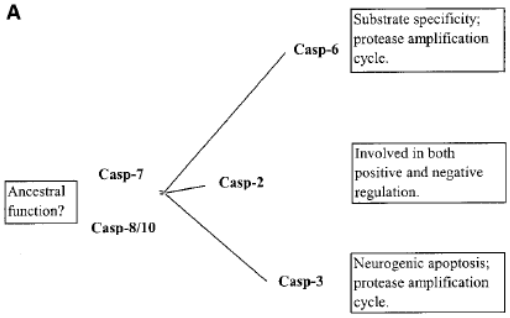
**TABLE 1**

**θ values and $d_F$ values from pairwise comparisons in the CED-3 subfamily**

| | ($\theta_{ij}$ ± SE) | | | | |
|---|---|---|---|---|---|
| $d_F(i, j)$ ± SE | Caspase-3 | Caspase-7 | Caspase-6 | Caspase-8/-10 | Caspase-2 |
| Caspase-3 | | 0.437 ± 0.178 | 0.844 ± 0.113 | 0.467 ± 0.113 | 0.540 ± 0.110 |
| Caspase-7 | 0.574 ± 0.257 | | 0.579 ± 0.198 | 0.006 ± 0.022 | 0.198 ± 0.184 |
| Caspase-6 | 1.858 ± 0.724 | 0.865 ± 0.470 | | 0.527 ± 0.190 | 0.627 ± 0.125 |
| Caspase-8/10 | 0.629 ± 0.212 | 0.006 ± 0.022 | 0.749 ± 0.401 | | 0.306 ± 0.180 |
| Caspase-2 | 0.777 ± 0.239 | 0.221 ± 0.229 | 0.986 ± 0.335 | 0.365 ± 0.259 | |

$\theta_{ij}$, the coefficient of the functional divergence between clusters $i$ and $j$; $d_F(i, j)$, the distance of the functional divergence between clusters $i$ and $j$; $d_F(i, j) = -\ln(1 - \theta_{ij})$, where $i$ and $j$ are the row and column designation numbers, respectively, and $i \neq j$; SE, standard error.

**A**




**B**

| cluster | $b_F$ | $d_N/d_S$ |
|---|---|---|
| Casp-3 | 0.694 | 0.085 |
| Casp-7 | -0.030 | 0.118 |
| Casp-6 | 0.901 | 0.071 |
| Casp-8/10 | -0.002 | 0.284 |
| Casp-2 | 0.198 | 0.086 |

FIGURE 6.—(A) A star-like topology of the CED-3 caspases in terms of type I functional branch length $b_F$. Biological evidence of functional specification for each caspase cluster is shown in the stacked boxes. (B) Functional branch length ($b_F$) and the ratio of nonsynonymous to synonymous rates ($d_N/d_S$) for each gene cluster, which were computed by using human-mouse sequences.

# Unit-3

# Functional Divergence after Gene Duplication: the Markov Chain Model

### Introduction

Since gene family proliferation is thought to have provided the raw materials for functional innovations, it is desirable, from sequence analysis, to identify amino acid sites that are responsible for the functional diversity. This approach has great potential for functional genomics because it is cost-effective, and these predictions can be further tested by experimentation. For molecular evolutionists, it is important to know the level of functional divergence after gene (or genome) duplication, as well as how many amino acid substitutions are actually involved in functional innovations. Since most amino acid changes are not related to functional divergence but represent neutral evolution, it is crucial to develop appropriate statistical methods to distinguish between these two possibilities.

When sequences of a gene family are available, the identification of functionally important residues can be approached computationally. The approach introduced by Casari et al. (1995) used a vectorial analysis of sequence profiles to identify functionally important residues. Lichtarge et al. (1996) developed a method called evolutionary tracing, which is extended recently by Landgraf et al. (1999), known as weighted evolutionary tracing. In these methods, the degree of conservation in each position is scored for different subfamilies and then visualized on the three-dimensional protein structure.

Gu (1999) has developed a probabilistic model, based on the underlying principle that functional divergence after gene duplication is highly correlated with the change of evolutionary rate. This correlation is complement to a fundamental rule in molecular evolution -- functional importance is highly correlated with evolutionary conservation (Kimura 1983). A site-specific profile based on posterior probability was then developed to predict critical residues for functional differences between two gene clusters. Wang and Gu (2000) have successfully applied this new-developed method to study the functional diversity of caspase gene family, and found that our predictions are supported by experimental data.

In this paper, the modeling for functional divergence after gene duplication is studied extensively under the Markov chain model of sequence evolution (Felsenstein 1981), which is further extended to the case of large family with many member genes. According to the observed alignment pattern (amino acid configuration), we study two important types of functional divergence (type I and type II, respectively). We show that Gu's (1999) method is a fast algorithm for two gene clusters. The performance of these methods is compared by examples.

### Functional divergence after gene duplication

Consider a multiple alignment of a gene family with two homologous genes 1 and 2 (figure 1). The pattern of amino acid alignment can be tentatively classified as follows (figure 1).
Type 0 represents the amino acid pattern that is universally conserved through the whole gene family, implying that those residues are important for the common function shared by all member genes.
Type I represents the amino acid pattern that is highly conserved in gene 1 but highly variable in gene 2, or *vice versa*, implying that those residues have experienced altered functional constraints.
Type II represents the amino acid pattern that is highly conserved in both genes but their biochemical properties are very different, e.g., charge positive v.s. negative, implying that those residues may be responsible for functional specification.
Type U represents the amino acid pattern at many residues that are not such clear-cut, referred as unclassified.

| | Sequence | Type 0 | Type 1 | Type 2 | Type-U |
|---|---|---|---|---|---|
| Gene 1 | 1 | CR | WQLV | RV | KTLI |
| | 2 | CR | WQIV | RV | RVLI |
| | 3 | CR | WQVG | RV | KIIV |
| | 4 | CR | WQVG | RV | NVLL |
| | 5 | CR | WQAT | RV | DMLL |
| | 6 | CR | WQAT | RV | IKIL |
| | 7 | CL | WQVI | RV | EKLI |
| | 8 | CR | WQIT | RV | DLVL |
| Gene 2 | 9 | CR | LTFD | DR | LKLM |
| | 10 | CR | ITFD | DR | QLLV |
| | 11 | CR | ITFD | ER | RLVV |
| | 12 | CR | YSFD | DK | LHVV |
| | 13 | CR | LEFD | DR | KMAL |
| | 14 | CL | LEFE | DR | KLLI |
| | 15 | CR | LEFD | DR | KLLL |
| | 16 | CR | VGFD | DK | ELII |
| | 17 | CR | VTFD | DR | RLII |

FIG. 1.—A hypothetical multiple alignment to show universally conserved sites (type 0), type I and type II amino acid configurations, and type U sites (unclassified).

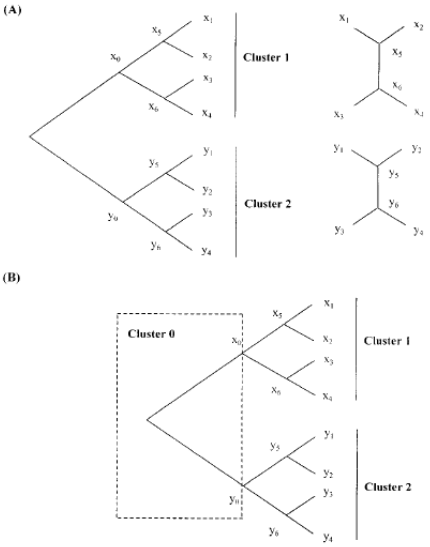



FIG. 2.—Gene clusters for the subtree likelihood (*A*) and for the whole-tree likelihood (*B*). The early stage after gene duplication is designated cluster 0.

**Table 1**
**Combined States (functional divergence configurations) for the Subtree Likelihood with Three Gene Clusters**

| State ($S_i$) | $F_0/F_1$ | $P(S_i)$ | Rate independence[a] | Type I Functional Divergence[b] |
|---|---|---|---|---|
| $S_0$ .... | $(F_0, F_0, F_0)$ | $\pi_0$ | $\lambda_1 = \lambda_2 = \lambda_3$ | No |
| $S_1$ .... | $(F_1, F_0, F_0)$ | $\pi_1$ | $\lambda_1, \lambda_2 = \lambda_3$ | Cluster 1 |
| $S_2$ .... | $(F_0, F_1, F_0)$ | $\pi_2$ | $\lambda_1 = \lambda_2, \lambda_3$ | Cluster 2 |
| $S_3$ .... | $(F_0, F_0, F_1)$ | $\pi_3$ | $\lambda_1, \lambda_2 = \lambda_3$ | Cluster 3 |
| $S_4$ .... | $(F_0, F_1, F_1)$ | $\pi_4$ | $\lambda_1, \lambda_2, \lambda_3$ | Clusters 2, 3 |
| | $(F_1, F_0, F_1)$ | $\pi_4$ | $\lambda_1, \lambda_2, \lambda_3$ | Clusters 1, 3 |
| | $(F_1, F_1, F_0)$ | $\pi_4$ | $\lambda_1, \lambda_2, \lambda_3$ | Clusters 2, 3 |
| | $(F_1, F_1, F_1)$ | $\pi_4$ | $\lambda_1, \lambda_2, \lambda_3$ | Clusters 1, 2, 3 |

[a] Rate independence under each state can be illustrated by the following example: $\lambda_1, \lambda_2 = \lambda_3$ means that $\lambda_1$ is independent of $\lambda_2$ or $\lambda_3$.
[b] Indicates which cluster(s) is under (type I) functional divergence.

**Table 2**
**Combined States (functional divergence configurations) for the Whole-Tree Likelihood with Two Gene Clusters**

| State ($S_i$) | $F_0/F_1$ | $P(S_i)$ | Rate independence[a] | Functional Divergence[b] |
|---|---|---|---|---|
| $S_0$ .... | $(F_0, F_0, F_0)$ | $\pi_0$ | $\lambda_1 = \lambda_2 = \lambda_3$ | No |
| $S_1$ .... | $(F_1, F_0, F_0)$ | $\pi_1$ | $\lambda_1, \lambda_2 = \lambda_3$ | Type II |
| $S_2$ .... | $(F_0, F_1, F_0)$ | $\pi_2$ | $\lambda_1 = \lambda_2, \lambda_3$ | Type I |
| $S_3$ .... | $(F_0, F_0, F_1)$ | $\pi_3$ | $\lambda_1 = \lambda_2, \lambda_3$ | Type I |
| $S_4$ .... | $(F_0, F_1, F_1)$ | $\pi_4$ | $\lambda_1, \lambda_2, \lambda_3$ | Type I |
| | $(F_1, F_0, F_1)$ | $\pi_4$ | $\lambda_1, \lambda_2, \lambda_3$ | Type I |
| | $(F_1, F_1, F_0)$ | $\pi_4$ | $\lambda_1, \lambda_2, \lambda_3$ | Type I |
| | $(F_1, F_1, F_1)$ | $\pi_4$ | $\lambda_1, \lambda_2, \lambda_3$ | Type I |

[a] Rate independence under each state can be illustrated by the following example: $\lambda_0, \lambda_1 = \lambda_2$ means that $\lambda_0$ is independent of $\lambda_1$ or $\lambda_2$.
[b] Type of functional divergence under each state.

After gene duplication, functional divergence between homologous genes 1 and 2 is likely to occur in the early stage (figure 2). According to amino acid configurations that are likely to be generated, there are two basic types of functional divergence after gene duplication. *Type I functional divergence* results in altered functional constraints (i.e., different evolutionary rate) between duplicate genes. *Type II functional divergence* results in no altered functional constraints but radical change in amino acid property between them (e.g., charge, hydrophobicity, etc.). Intuitively, one may expect that type I (or type II) amino acid patterns are likely to be generated by type I (or type II) functional divergence. It might be true only when the effect of type I (or type II) functional divergence has been shown statistically significant under a stochastic model. Then, the possibility of a site being functional divergence-related (type I or type II) can be measured by a posterior probability, when the observed amino acid pattern is given.

### Type I functional divergence: two clusters

Following the statistical framework developed by Gu (1999), we build a ``subtree'' likelihood to estimate the (type I) functional divergence by detecting the level of altered functional constraints between two clusters (figure 2A). The advantage of subtree likelihood is its simplicity because the phylogenetic relationship among gene clusters will not be considered.

*Unrooted likelihood for each cluster*

Under the Markov chain model, the likelihood for sequence evolution can be derived as follows. First, the transition probability matrix for a given time period $t$ can be computed as $\boldsymbol{P}=e^{\lambda \boldsymbol{R} t}$, where the rate matrix $\boldsymbol{R}$ represents the pattern of amino acid substitutions, which can be empirically determined by, for example, the Dayhoff model the JTT model. The evolutionary rate ($\lambda$) may vary among sites because of different functional constraints. Usually $\lambda$ is treated as a random variable, which follows a gamma distribution, i.e.,

$$\phi(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} \quad (1)$$

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), whereas $\beta$ is a scale constant. Consider the phylogenetic tree in Figure 2. Let $X=(x_1, x_2, x_3, x_4)$ and $Y=(y_1, y_2, y_3, y_4)$ be the observed amino acid patterns of a site for clusters 1 and 2, respectively. For the unrooted subtree for cluster 1 or 2 (panel A), the conditional probability of observing $X$ or $Y$ at a site can be written as follows, respectively

$$\begin{aligned} f(X|\lambda) &= \sum_{x_5=1}^{20} \sum_{x_6=1}^{20} b_{x_5} P_{x_5 x_1} P_{x_5 x_2} P_{x_5 x_6} P_{x_6 x_3} P_{x_6 x_4} \\ f(Y|\lambda) &= \sum_{y_5=1}^{20} \sum_{y_6=1}^{20} b_{y_5} P_{y_5 y_1} P_{y_5 y_2} P_{y_5 y_6} P_{y_6 y_3} P_{y_6 y_4} \end{aligned} \quad (2)$$

where $P_{ij}=P_{ij}(v_{ij})$ is the transition probability from node $i$ to node $j$, $v_{ij}$ is the branch length between them; $b_i$ is the frequency of amino acid $i$. By integrating out the random variable $\lambda$, the probability of observing $X$ or $Y$ at a site is given by

$$\begin{aligned} p(X) &= E[f(X|\lambda)] = \int_0^{\infty} f(X|\lambda)\phi(\lambda)d\lambda \\ p(Y) &= E[f(Y|\lambda)] = \int_0^{\infty} f(Y|\lambda)\phi(\lambda)d\lambda \end{aligned} \quad (3)$$

respectively, where $E$ means taking expectation.

*Two-state model for functional divergence*

For two gene clusters generated by gene duplication, the two-state model assumes that, in each cluster one site has two possible states, $F_0$ (functional divergence-unrelated) and $F_1$ (functional divergence-related). As a result, there are four combined states, i.e., ($F_0$, $F_0$), ($F_0$, $F_1$), ($F_1$, $F_0$), and ($F_1$, $F_1$). These states are also called *functional divergence configuration*, where the first position is for clusters 1 and the second for cluster 2. When a site is under ($F_0$, $F_0$), i.e., no altered functional constraints in both clusters, the evolutionary rate at this site is virtually the same between two clusters, i.e., $\lambda_1=\lambda_2$. For the last three combined states, however, the amino acid residue has experienced altered functional constraints (i.e., under $F_1$) at least in one cluster, resulting in statistical independence between $\lambda_1$ and $\lambda_2$ (Gu 1999).

The assumption of rate-independence for type I functional divergence means that knowing the evolutionary rate (or the functional constraint) at such sites in one cluster contains no information for predicting the intensity of functional constraint in the other cluster. Since $\lambda_1$ and $\lambda_2$ are independent under each of ($F_0$, $F_1$), ($F_1$, $F_0$) and ($F_1$, $F_1$), these combined states are not distinguishable under the current model; they have to be degenerated to a single one. Consequently, there are two nondegenerate combined states (functional divergence configurations), denoted by $S_0=(F_0, F_0)$, and $S_1=(F_0, F_1)+(F_1, F_0)+(F_1, F_1)$, respectively. It should be noted that ($F_0$, $F_0$) was written as $\boldsymbol{F_0}$, and $S_1$ as $\boldsymbol{F_1}$ (Gu 1999). In this sense, the *F*-notation describes the status in a single cluster, while *S*-notation is used for the functional divergence configuration of a gene family.

*The subtree likelihood*

Let $P(S_1)=\theta_{12}$ be the probability of a site being in state $S_1$, and $P(S_0)=1-\theta_{12}$ be the probability of a site being in state $S_0$. We call $\theta_{12}$ *the coefficient of type I functional divergence* between cluster 1 and cluster 2 (Gu 1999). Let $X$ and $Y$ be the amino acid pattern of a site in clusters 1 and 2, respectively. Our purpose is to build a likelihood function for estimating $\theta_{12}$ from sequences. Gu (1999) has shown that the subtree likelihood provides a simple solution for this purpose. Since it only depends on the (unrooted) subtrees of two clusters, the joint probability can be easily derived based on the pattern of rate-independence. In the following, we use the superscript (*) to distinguish the subtree likelihood from the conventional (whole-tree) likelihood, e.g., the joint probability of two subtrees is denoted by

$p^*(X, Y)$.

Since evolutionary rates ($\lambda_1$ and $\lambda_2$) at an $S_1$-site (i.e., a site under $S_1$), are statistically independent between two clusters, whereas they are completely correlated ($\lambda_1=\lambda_2$, without loss of generality) at an $S_0$-site, the joint probability of subtrees conditional on $S_0$ or $S_1$ is given by

$$\begin{aligned} f^*(X,Y|S_0) &= \int_0^\infty f(X|\lambda)f(Y|\lambda)\phi(\lambda)d\lambda = E[f(X|\lambda)f(X|\lambda)] \\ f^*(X,Y|S_1) &= p(X)p(Y) = E[f(X|\lambda_1)] \times E[f(Y|\lambda_2)] \end{aligned} \quad (4)$$

where $f(X|\lambda_1)$ or $f(Y|\lambda_2)$ is the likelihood for each unrooted subtree, respectively, e.g., it is given by Eq.(2) for the phylogeny in figure 2(A).

From the two-state model, one can easily show that the joint probability of two subtrees can be written as

$$p^*(X,Y) = (1-\theta_{12})f^*(X,Y|S_0) + \theta_{12}f^*(X,Y|S_1) \quad (5)$$

Then, under the assumption of site-independence, the likelihood function over all sites (gaps excluded) is given by

$$L^*(\mathbf{x}|data) = \prod_k p^*(X^{(k)}, Y^{(k)}) \quad (6)$$

where $k$ runs for sites, and $\boldsymbol{x}$ is the set of unknown parameters.

*Numerical algorithm*

It is complicated to compute $p^*(X, Y)$; they involve the phylogenetic tree, branch lengths ($\boldsymbol{v}$), the shape parameter (α) of a gamma distribution, and the coefficient of functional divergence ($\theta_{12}$). We propose the following algorithm to solve this problem:

(1) The phylogenetic tree is inferred by the neighbor-joining method (Saitou and Nei 1987), which can handle very large number of sequences.

(2) Given the inferred topology, the branch lengths ($\boldsymbol{v}$) are estimated by a least-square method, and the gamma shape parameter (α) is estimated by Gu and Zhang's (1997) method. Then, computation of those expectations in Eq.(4) can be approximated similar to Yang (1994).

(3) Regarding all other parameters as constants, the maximum likelihood estimate (MLE) of $\theta_{12}$ can be obtained by

$$\partial \ln L^* / \partial \ln \theta_{12} = 0$$

which satisfies

$$\sum_{k=1}^{N} \frac{1}{h_k + \hat{\theta}_{12}} = 0 \quad (7)$$

where $N$ is the sequence length and $h_k=1/(a_k-1)$, $a_k=f^*(X, Y|S_1)/f^*(X, Y|S_0)$ for site $k$.

(4) A numerical iteration such as simplex method is implemented to find the final ML estimates of $\boldsymbol{v}$, α and $\theta_{12}$ under the given phylogeny.

After obtaining these ML estimates, the likelihood ratio test (LRT) can be constructed under the null hypothesis $H_0$: $\theta_{12}=0$ vs. $H_A$: $\theta_{12}>0$. If $H_0$ is rejected significantly, it provides statistical evidence for functional divergence in the coding region after gene duplication, i.e., functional constraints have shifted between two homologous genes.

## Type I functional divergence: multiple clusters

For a large gene family with many member genes (clusters), the pattern of amino acid alignments is complicated. Even for three clusters, type I amino acid pattern contains many subtypes: cluster 1 is highly variable but clusters 2 and 3 are conserved, etc. The subtree likelihood can be extended to any $n$ number of gene clusters, but may demand a huge computational time when $n$ is large.

*Subtree likelihood for multiple clusters*

Let $\lambda_i$ and $X_i$ be the evolutionary rate and amino acid pattern of a site in cluster $i$, respectively, $i$=1,…,$n$, and X=($X_1$,…, $X_n$). For two possible states ($F_0/F_1$) in each gene cluster, we have in total $2^n$ possible combined states (functional divergence configurations).

For three gene clusters ($n$=3), all functional divergence configurations ($2^3$=8) are listed as follows: ($F_0$, $F_0$, $F_0$), ($F_0$, $F_0$, $F_1$), ($F_0$, $F_1$, $F_0$), ($F_1$, $F_0$, $F_0$), ($F_0$, $F_1$, $F_1$), ($F_1$, $F_0$, $F_1$), ($F_1$, $F_1$, $F_0$), ($F_1$, $F_1$, $F_1$), where the first, second and third positions refer to the $F_0/F_1$ status of gene clusters 1, 2, and 3, respectively. For each of them, the relationship of

evolutionary rates ($\lambda_1$, $\lambda_2$ and $\lambda_3$) among three clusters is shown in table 1. If a site is under ($F_0$, $F_0$, $F_0$), $\lambda_1$, $\lambda_2$ and $\lambda_3$ are completely correlated so that one can assume $\lambda_1=\lambda_2=\lambda_3$, without loss of generality. Similarly, under ($F_0$, $F_0$, $F_1$), $\lambda_1=\lambda_2$ ($=\lambda$), which is independent of $\lambda_3$; and so forth. However, the last four combined states, ($F_0$, $F_1$, $F_1$), ($F_1$, $F_0$, $F_1$), ($F_1$, $F_1$, $F_0$), ($F_1$, $F_1$, $F_1$), have to be degenerated to a single combined state ($S_4$) because $\lambda_1$, $\lambda_2$ and $\lambda_3$ are mutually independent (table 1). Note that there are $m=2^n-n$ nondegenerate combined states (functional divergence configurations) in the case of $n$ clusters, which are denoted by $S_j$, $j=0,\ldots, m-1$. In particular, $S_0$ indicates the functional divergence configuration that all clusters are under $F_0$, i.e., $S_0=(F_0, F_0, \ldots, F_0)$. According to the pattern of rate independence (table 1), we can show that the joint probability of three subtrees under each functional divergence configuration ($S_j$) is given by

$$\begin{aligned} f^*(X|S_0) &= E[f(X_1|\lambda)f(X_2|\lambda)f(X_3|\lambda)] \\ f^*(X|S_1) &= E[f(X_1|\lambda_1)] \times E[f(X_2|\lambda)f(X_3)|\lambda)] \\ f^*(X|S_2) &= E[f(X_2|\lambda_2)] \times E[f(X_1|\lambda)f(X_3|\lambda)] \\ f^*(X|S_3) &= E[f(X_3|\lambda_3)] \times E[f(X_1|\lambda)f(X_2|\lambda)] \\ f^*(X|S_4) &= E[f(X_1|\lambda_1)] \times E[f(X_2|\lambda_2)] \times E[f(X_3|\lambda_3)] \end{aligned} \quad (8)$$

where $f(X_1|\lambda)$, $f(X_2|\lambda)$ or $f(X_3|\lambda)$ is the likelihood for the (unrooted) subtree of each gene cluster, respectively. Let $\pi_j$ be the (prior) probability of a site under $S_j$, i.e., $\pi_j=P(S_j)$. Thus, the joint probability of three subtrees at a site is given by

$$p^*(X) = \sum_{j=0}^{m-1} \pi_j f^*(X|S_j) \quad (9)$$

where $m=5$. Apparently, Eq.(5) is a special case of Eq.(9) when $n=2$ (and so $m=2$), and $\pi_0=1-\theta_{12}$ and $\pi_1=\theta_{12}$. In general, $\pi_j$ ($j=1, 2,\ldots, m-1$) is called the coefficients of type I functional divergence for functional divergence configuration $S_j$. In particular, we define

$$\pi_f = 1 - \pi_0 = \sum_{j=1}^{m-1} \pi_j \quad (10)$$

as the coefficient of (type I) functional divergence of the gene family.

*Numerical algorithm*

Similar to Eq.(6), unknown parameters can be estimated by maximizing the likelihood $L^*=\Pi_k p^*(X^{(k)})$, which can be approached by extending the algorithm for two-clusters; they are the same except for step 3 [i.e., Eq.(7)]. When the number of gene clusters ($n$) is not very large, the Newton-Raphson algorithm is computationally efficient. Let $\boldsymbol{\pi}$ be the parameter vector, $\boldsymbol{\pi}=(\pi_1,\ldots, \pi_{m-1})$. The iteration equation is then given by $\boldsymbol{\pi}^{(l+1)}= \boldsymbol{\pi}^{(1)}-\boldsymbol{H}^{-1}\boldsymbol{g}$, where $\boldsymbol{g}$ is the grade vector, whose $i$-th element is

$$g_i = \partial \ln L^* / \partial \pi_i$$

and $\boldsymbol{H}$ is the Haesson matrix, whose $ij$-th element is

$$H_{ij} = \partial^2 \ln L^* / \partial \pi_i \partial \pi_j$$

When appropriate initial values are given, $\boldsymbol{\pi}^{(1)}$ would converge to $\boldsymbol{\pi}$^. Finally, their large sample variance-covariance matrix can be approximately estimated by the inverse of Fisher's information matrix.

When $n$ is large, an EM algorithm (Expectation and Maximization) can be implemented. The EM method is a very general iterative approach for the dataset with missing (or incomplete) data. In our case, the ML estimates of $\pi_i$ would be easy to obtain, if we know the state ($F_0/F_1$) to which each site belongs in each gene cluster. Thus, the original data set is treated as incomplete data, missing the category information. Using a current estimate of the unknown parameter values, the expected value of the incomplete data is computed, weighted by the posterior probability. This is the expectation, or **E**-step. The result is a set of likelihood equations that are considerably easier to solve than the full likelihood (the maximization, or **M**-step). The new estimates obtained from the M step are then used to update the expected values, and this approach is iterated until convergence.

Likelihood ratio tests (LRT) can be constructed under various null hypothesis by specifying some coefficients of functional divergence. In particular, the LRT under the null $H_0$: $\pi_1=1$ is apparently the most powerful test.

Significant rejection of the null suggests functional divergence among member genes of a gene family.

### Types I and II functional divergences: two-gene clusters

In spite of the efficiency for estimating type I functional divergence, the subtree likelihood is not able to detect type II functional divergence that requires the evolutionary relationship between clusters. Therefore, it is desirable to build a 'whole-tree' likelihood for estimating these two types of functional divergence simultaneously.

In the early stage after gene duplication, many evolutionary forces (e.g., positive selection, functional relaxation, or coevolution between contact sites) may play roles in amino acid substitutions so that a comprehensive modeling could be complicated. A simple solution is to consider the internal branch between two clusters (i.e., the early stage) as cluster 0 (figure 2B), which is ancestral. Let $\lambda_1$ and $\lambda_2$ be the evolutionary rates in clusters 1 and 2, respectively, and $\lambda_0$ be the evolutionary rate in the internal branch (cluster 0). For each cluster, a given site has two possible states, $F_0$ (functional divergence-unrelated) and $F_1$ (functional divergence-related). Therefore, similar to the subtree likelihood of three clusters, we have $2^3$=8 possible combined states that can be degenerated to 5 functional divergence configurations, under which the relationship between $\lambda_0$, $\lambda_1$ and $\lambda_2$ is shown in Table 2. Let $\pi_j$ ($j$=1,…,5) be the probability of a site being under $S_j$, i.e., $\pi_j=P(S_j)$. For a given site, the conditional probability for observing $X$ and $Y$ is given by

$$f(X,Y|\lambda)=\sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0}P_{x_0y_0}(v|\lambda_0)f(X|\lambda_1;x_0)f(Y|\lambda_2;y_0) \quad (11)$$

where $f(X|\lambda_1; x_0)$ and $f(Y|\lambda_2; y_0)$ are the likelihood functions for clusters 1 and 2, conditional of the roots $x_0$ and $y_0$, respectively, and $\boldsymbol{v}$ is the internal branch length. When the phylogeny is given in figure 2B, we have

$$\begin{aligned} f(X|\lambda;x_0) &= \sum_{x_5}\sum_{x_6} P_{x_0x_5}P_{x_5x_1}P_{x_5x_2}P_{x_0x_6}P_{x_6x_3}P_{x_6x_4} \\ f(Y|\lambda;y_0) &= \sum_{y_5}\sum_{y_6} P_{y_0y_5}P_{y_5y_1}P_{y_5y_2}P_{y_0y_6}P_{y_6y_3}P_{y_6y_4} \end{aligned} \quad (12)$$

The conditional probability for observing X and Y under each combined state is given by

$$\begin{aligned} f(X,Y|S_0) &= \sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0}E[P_{x_0y_0}(v|\lambda_0)f(X|\lambda;x_0)f(Y|\lambda;y_0)] \\ f(X,Y|S_1) &= \sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0}E[P_{x_0y_0}(v|\lambda_0)]\times E[f(X|\lambda;x_0)f(Y|\lambda;y_0)] \\ f(X,Y|S_2) &= \sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0}E[f(X|\lambda_1;x_0)]\times E[P_{x_0y_0}(v|\lambda_0)f(Y|\lambda;y_0)] \\ f(X,Y|S_3) &= \sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0}E[P_{x_0y_0}(v|\lambda_0)f(X|\lambda;x_0)]\times E[f(Y|\lambda_2;y_0)] \\ f(X,Y|S_4) &= \sum_{x_0=1}^{20}\sum_{y_0=1}^{20} b_{x_0}E[P_{x_0y_0}(v|\lambda_0)]\times E[f(X|\lambda_1;x_0)]\times E[f(Y|\lambda_2;y_0)] \end{aligned} \quad (13)$$

Therefore, the joint probability of X and Y can be generally expressed as follows

$$p(X,Y)=\sum_{j=0}^{m-1}\pi_j f(X,Y|S_j) \quad (14)$$

where $m$=5 in this case. Similar to above, maximization of the likelihood $L=\Pi_k\, p(X^{(k)}, Y^{(k)})$ can be achieved by either Newton-Raphson or EM algorithm.

Next we show how Eqs.(13) and (14) are related to calculate the coefficients of type I and type II functional divergence. Since type II functional divergence results no altered functional constraints between two clusters, it can be interpreted as the functional divergence configuration $S_1=(F_1, F_0, F_0)$, i.e., cluster 0 is under $F_1$, but clusters 1 and 2 are under $F_0$. Therefore, the coefficient of type II functional divergence can be defined as $\theta_{II}=P(S_1)=P(F_1, F_0, F_0)=\pi_1$. On the other hand, type I functional divergence means that at least either cluster 1 or cluster 2 should be under $F_1$, regardless of the status of cluster 0. According to table 2, the coefficient of type I functional divergence is given by $\theta_I=P(S_2)+P(S_3)+P(S_4)$. Moreover, if the coefficient of overall functional divergence is defined as $\pi_f=1-P(S_0)=1-\pi_0$, we have

$$\theta_I+\theta_{II}=\theta_f=1-\pi_0 \quad (15)$$

Thus, $\pi_0$ can be called the coefficient of functional constraint of the gene family.

## Predicting critical residues for functional divergence

It is of great interest to (statistically) predict which sites are likely to be responsible for these type I and type II functional differences. Indeed, these sites can be further tested by experimentation, using molecular, biochemical or transgenic approaches. We will develop *site-specific profiles* for this purpose, which can be achieved by the posterior analysis.

*Type I functional divergence predicted from the subtree likelihood*

For the simple two-cluster case, there are only two nondegenerate states: $S_0$ and $S_1$. We wish to know the probability of $S_1$ for a given site when the amino acid configuration ($X$, $Y$) is observed, i.e., $P^*(S_1|X, Y)$. The prior probability of S1 is $P(S1)= \pi_1=\theta_1$. According to the Bayesian law, we have

$$P^*(S_1|X,Y) = \frac{\theta_I f^*(X,Y|S_1)}{p^*(X,Y)} \quad (16)$$

where $f^*(X, Y|S_1)$ and $p^*(X, Y)$ are given by Eqs.(4) and (5), respectively.

Now we consider the case of multiple clusters. Similarly, when the amino acid configuration (X) at a site is given, the posterior probability of each functional divergence configuration $S_i$ can be generally expressed as follows

$$P^*(S_i|X) = \frac{\pi_i f^*(X|S_i)}{\sum_{j=0}^{m-1} \pi_j f^*(X|S_j)}, \quad i = 0, 1, \ldots, m-1 \quad (17)$$

Where $\pi_i=P(S_i)$ is the prior probability of state $S_i$. When $n$=2, Eq.(17) is reduced to Eq.(16).

For a large gene family with little knowledge about its functional diversity, a site-specific measure for the overall level of type I functional divergence at each site is useful. Since the coefficient of overall functional divergence of a gene family is defined as $\pi_f=1-\pi_0$, where $\pi_0=P(S_0)$ for $S_0=(F_0,\ldots,F_0)$, it is natural to define such measure as $P^*(F_1|X)=1-P^*(S_0|X)$. According to Eq.(17), it is given by

$$P^*(\mathbf{F}_1|X) = 1 - \frac{\pi_0 f^*(X|S_0)}{\sum_{j=0}^{m-1} \pi_j f^*(X|S_j)} \quad (18)$$

*Type I and type II functional divergence*

Based on the whole-tree likelihood for functional divergence, we can develop a site-specific profile for type I as well as type II functional divergence. In the case of two-clusters, the posterior probability of each (nondegenerate) combined state $S_i$ (table 2) can be computed as

$$P(S_i|X,Y) = \frac{\pi_i f(X,Y|S_i)}{\sum_{j=0}^{m-1} \pi_j f(X,Y|S_j)}, \quad i = 0, 1, \ldots, 4 \quad (19)$$

where $\pi_i=P(S_i)$, and $f(X, Y|S_j)$ is given by Eq.(13). Thus, one can easily show that site-specific profiles for type I and type II functional divergence are given by

$$\begin{aligned} P(\text{type I}|X,Y) &= P(S_2|X,Y) + P(S_3|X,Y) + P(S_4|X,Y) \\ P(\text{type II}|X,Y) &= P(S_1|X,Y) \end{aligned} \quad (20)$$

respectively.

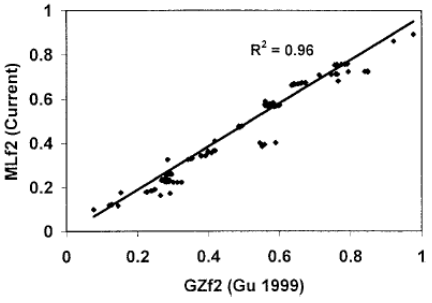



FIG. 7.—The correlation of two site-specific profiles (posterior probabilities), which are computed by the current method and Gu's (1999) method.

# Unit-4

# Type-II Functional Divergence after Gene Duplication

## Introduction

According to the theory of molecular evolution (Kimura 1983), 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).

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).

In this study, we develop a statistical method for type-II functional divergence. To this end, we have to address two related statistical issues. First, are the type-II 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?

## Modeling type-II functional divergence in the early stage

*The two-state model*

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; 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 $\theta_{II}$ *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 $F_0$*

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, or the JTT model. 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.** 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

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

respectively, where $t$ is the evolutionary time, $\lambda$ is the substitution rate, and $\pi_R$ (or $\pi_C$) is the proportion of radical (or conserved) substitutions in the total substitutions; $\pi_R+\pi_C=1$. Apparently, 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).

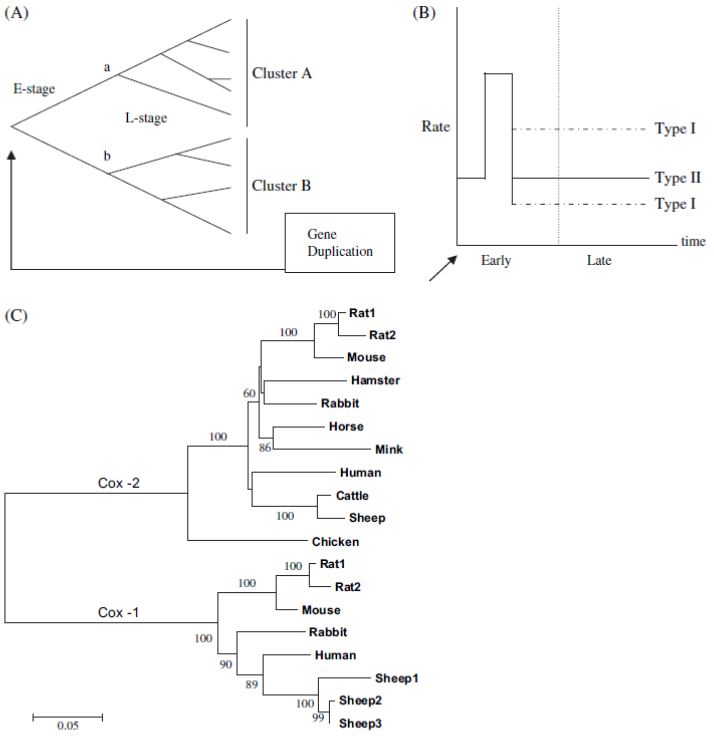



FIG. 1.—(*A*) Two gene clusters after gene duplication. E and L are early and late stages of gene cluster A and B, respectively. (*B*) Type-I functional divergences: in the early stage, the evolutionary rate may increase for functional divergence–related change, resulting in shifted functional constraints 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 a 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 of more than 50% are presented. Modified from Gu (1999, 2001).

*Substitution models under $F_1$*

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 $P(N|F_1)=0$. Together, one may write

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

That is, $a_R$ (or $a_C$) is the ($F_1$)-proportion of radical (or conserved) substitutions in total substitutions. Moreover, the

$F_1$-radical amino acid substitution ($a_R$) can be much higher than that under $F_0$ ($\pi_R$), as will be shown later.

### Evolutionary link between early and late stages

The evolutionary link between early and late stages depends on the status of type-II functional divergence. Let $\lambda_E$ and $\lambda_L$ 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

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

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 α=∞ means no rate variation among sites.

(*ii*) Under $F_0$, the evolutionary rates in the early ($\lambda_E$) and late ($\lambda_L$) stages share the same rate component *u*. That is, $\lambda_E=c_1u$ and $\lambda_L=c_2u$, where $c_1$ and $c_2$ are constants.

(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, and $P(X|Y)$ 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). 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

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

Direct application of Eq.(4) for estimating $\theta_{II}$ 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 a genome-wide analysis.

### 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

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

with the same applying to $p_B(k)$, where $T_A$ (or $T_B$) is the total evolutionary time of cluster *A* (or *B*), and $\lambda_A$ and $\lambda_B$ 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, $P(X|Y)=p_Ap_B$, which is independent of the early stage $Y$. Similar to the derivation of Eq.(4), we have

$$\begin{aligned} P(X=(i,j),Y|F_0) &= \int_0^\infty P(Y|F_0)p_A(i)p_B(j)\phi(u)du \\ P(X=(i,j),Y|F_1) &= \int_0^\infty P(Y|F_1)p_A(i)p_B(j)\phi(u)du \end{aligned}$$

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 $Z_B=D_B/(D_A+D_B+\alpha)$; $D_A=E[\lambda_A]T_A$ and $D_B=E[\lambda_B]T_B$ be the total branch lengths of clusters $A$ and $B$, respectively, and $\alpha$ is the gamma shape parameter.

Next we consider the early-stage distribution $f_Y$, the frequencies of three early-stage amino acid patterns for $Y=N$, $R$ or $C$. Since $f_Y=\Sigma_i f_{ij,Y}$, 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 $d=E[\lambda]t$ 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.

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

where $M_{ij}$ is given by

$$\begin{aligned} M_{ij} &= \int_0^\infty e^{-\lambda t} p_A(i)p_B(j)\phi(u)du \\ &= \frac{\Gamma(i+j+\alpha)}{i!\ j!\ \Gamma(\alpha)} W^\alpha W_A^i W_B^j \end{aligned} \quad (11)$$

### Estimation procedure

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 $\alpha$. We thus modified the likelihood method of Gu and Zhang (1997) to estimate them simultaneously, $D_A$^, $D_B$^ and α, 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 $Y=N$, $Y$ or $C$. After treating three late-stage parameters as known, we develop a simple likelihood to estimate early-stage parameters $\theta_{II}$, $a_R/a_C$, 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

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

includes two unknown parameters $\theta_{II}$ and $d$. Let $N_0=\Sigma_{i,j}\, n_{ij,N}$ be the total number of sites that have no change in the early stage. Under the $p$-constraint of Eq.(9), the ML estimate of $\theta_{II}$ is given by $\theta_{II}\hat{}=1/(1-y)$, where $y$ is the solution of

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

with $d=-\ln(1-p)+\ln(1-\theta_{II})$. (Note that $M_{ij}$ depends on the parameter $d$, while $Q_{ij}$ only depends on 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, $f_{ij,S}\hat{}=n_{ij,S}/L$ and $f_0\hat{}=N_0/L$. The sampling variance of $\theta_{II}\hat{}$a can be calculated as follows

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

where $a=\Sigma_{i,j}\, f_{ij,S}\, M_{ij}^2/(Q_{ij}-M_{ij}+M_{ij}\theta_{II}^2$. When the estimates of $\theta_{II}$ and $d$ 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 (Yang et al 1995) to infer the ancestral sequences of $Y$. Then we estimate $p$ when each site in the inferred ancestral sequence receives the assignment of amino acid with the highest posterior probability.

*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

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

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

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

The sampling variance of the estimate $\theta_{II}\hat{}$ is $\mathrm{Var}(\theta_{II})=f_{00N}(1-f_{00N})b^2/N$, where $b=[(1+D_A+D_B+d)/\alpha]^{\alpha}$. 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 functional differences is of great interest, if the coefficient of functional divergence ($\theta_{II}$) 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

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

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 $\mathrm{P}(X=(i, j), Y)$ 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 F1 to F0, 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 $h=d/(D_A+D_B+d+\alpha)$.

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 $Y \geq N$. 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 $\theta_{II}$ is not significantly larger than 0. Finally, we note that $R(F1|F0)max \to \infty$ 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.

**Table 1**
**Summary of Functional-Divergence Analysis for COX and G-Protein Alpha Families**

| | COX | G-Protein Alpha |
|---|---|---|
| Type II | | |
| *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 |
| $a_R/\pi R$ | 2.744 | 2.811 |
| $\theta_{II}$ + SE | 0.159 ± 0.036 | 0.325 ± 0.055 |
| Type I | | |
| $\theta_I$ + SE | 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 fig. 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 $D_b$ are the average numbers of substitutions per sites in clusters A and B, respectively, and *d* is the distance between nodes *a* and *b*. The parameter α is the gamma shape parameter. $f_R$ is the observed proportion of radical changes in all substitutions between nodes *a* and *b*. $a_R/\pi_R$ is the ratio of radical changes under (type-II) functional divergence versus nonfunctional divergence. Finally, $\theta_I$ and $\theta_{II}$ are the coefficients of type-I and type-II functional divergence, respectively. SE: standard error.

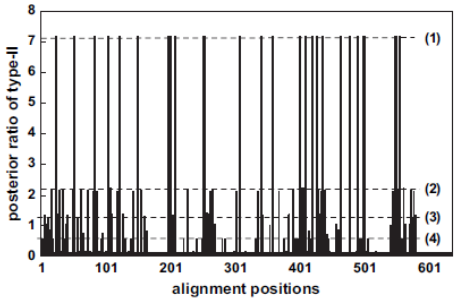



FIG. 2.—Site-specific profile for type-II functional divergence between COX1 and COX2, measured by the posterior ratio. Horizontal lines (1)–(4) indicate cluster-specific patterns in table 2.

**Table 2**
**Functional Ranking of Several Cluster-Specific Patterns in the COX Gene Family**

| | Between Clusters (early stage) | Within Clusters (late stage) | Number of Sites | Ratio Score | Posterior Probability |
|---|---|---|---|---|---|
| (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 |

NOTE.—Pattern (1): radical cluster-specific sites. Patterns (2)–(3): imperfect radical cluster-specific sites. Pattern (4): conserved cluster-specific sites. a.a.: amino acid.

**Table 3**
**Summary of Amino Acid Changes in 22 Radical Cluster-Specific Positions Associated with the Divergence of COX1 and COX2**

| Position | COX1 | COX2 | Property Change |
|---|---|---|---|
| 22 | Y | S | H versus P0 |
| 51 | P | E | P0 versus − |
| 82 | W | G | H versus P0 |
| 103 | V | S | H versus P0 |
| 121 | I | K | H versus + |
| 149 | T | V | P0 versus H |
| 197 | S | D | P0 versus − |
| 251 | E | K | − versus + |
| 253 | A | T | H versus P0 |
| 306 | T | E | P0 versus − |
| 340 | F | H | H versus + |
| 358 | R | Q | + versus P0 |
| 401 | Y | H | H versus + |
| 409 | A | S | H versus P0 |
| 419 | G | A | P0 versus H |
| 425 | D | P | − versus P0 |
| 427 | H | A | + versus H |
| 435 | V | S | H versus P0 |
| 463 | Q | E | P0 versus − |
| 499 | S | A | P0 versus H |
| 548 | K | Q | + versus P0 |
| 555 | T | V | P0 versus H |

NOTE.—H: hydrophobic; P0: hydrophilic with neutral charge; +: charge positive; and −: charge negative.

# Unit-5

# Type-II Functional Divergence after Gene Duplication: GPCR Gene Family Analysis

## Introduction

*Drug-design and side effects*

Precision medicine enables the development of targeted drugs and improvement of the therapeutic efficacy [1]. However, some targeted drugs are promiscuous, showing a high risk of severe side effects because they have unexpected targets and exhibit low specificity [2]. Cross-reactivity on protein paralogs may cause undesirable side effects of drugs [3]. Generated from gene duplications, paralogs are evolutionally homologous [4] and share similar protein sequences or structural features, thus comprising similar binding pockets with drugs. As a result, a drug that binds to the target protein encoded by one gene may also bind to its paralog [5].

Because most drug targets are paralogs [3], controlling target specificity to minimize side effects is required to create novel and safer drugs. Such control may be achieved by drug design guided by paralog-discriminating features, known as ‘‘selectivity filters” [3]. Therefore, identifying evolutionally divergent features that enable paralog discrimination would be beneficial. It is well accepted that amino acids are evolution- ally conserved if they are functionally important [6]. Therefore, an amino acid residue is said to be functionally or structurally important if it is evolutionally conserved [7], whereas an evolutionally-variable residue is said to be less important. It is thus believed that alterations in the evolutionary conservation at a particular residue imply that this residue may have been involved in the functional divergence of a gene family during the evolution [4].

*Functional divergence*

Type-I functional divergence gives rise to the site-specific rate variation after gene duplication [8,9]. Typically, an amino acid residue related to the type-I functional divergence is highly conserved in one duplicate gene, but highly variable in the other one. Drug binding sites tend to be functionally important. If a drug targets the conserved residue of type-I functionally-divergent site in one paralog, its binding to the non-conserved residue in another paralogs would be avoided. Therefore, the alteration in evolutionary conservation resulting from type-I functional divergence can distinguish one paralog from another, which may reduce the occurrence of cross-reactivity.

Type-II functional divergence brings about the change of site-specific property. Typically, amino acid residues are highly evolutionally conserved within each cluster of orthologous genes, *i.e.*, both residues play vital roles functionally or structurally for this gene family. However, a radical change of amino acid property at a homologous site occurred between the two duplicate genes. For example, one residue is positively- charged in a gene but its homologous residue in the duplicated gene is negatively-charged [10,11]. If a drug is designed to be negatively-charged, it can bind to a positively-charged residue in one paralog, but not the negatively-charged one in another paralog. A shift in key physicochemical properties relevant to ligand binding interactions may result in alterations in binding features or affect the druggability of protein targets [12]. Therefore, type-II functional divergence features in physicochemical properties between paralogs can be exploited as selectivity filters to function as targetable differences [13].

*G-protein coupled receptors (GPCRs)*

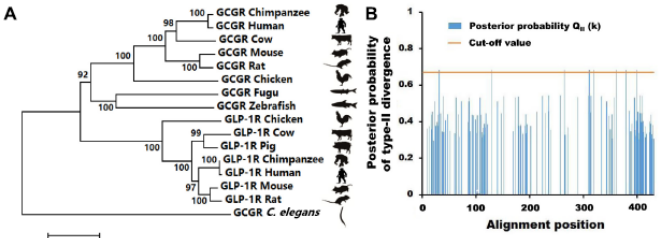
The known target protein receptor family of G-protein cou- pled receptors (GPCRs) contributes significantly to side effects [14]. GPCRs constitute one of the largest families of membrane proteins with approximately 800 members encoded in the human genome [15]. According to the GRAFS classification system, GPCRs fall into five categories, including glutamate (G), rhodopsin (α, β, γ and δ) (R), adhesion (A), frizzled/- taste2 (F), and secretin (S) families [16]. It is estimated that 30%–40% of all drugs currently on the market target GPCRs [17]. Since the gene members of this superfamily arose from gene duplication [18], these gene targets are rich in paralogs.

## Functional role difference between paralog GCGR and GLP-1R

This study aimed to reduce the side effects caused by paralogs. Glucagon receptor (GCGR) and glucagon-like

peptide-1 receptor (GLP-1R), two clinically validated drug targets in patients with type 2 diabetes, were used as an example. The glucagon-like subfamily belongs to secretin type GPCRs and is rich in clinically validated targets [18]. This subfamily constitutes 4 hormone receptors duplicated from the early stage of vertebrates [19]. These receptors play crucial roles in hormonal homeostasis in humans and other animals and serve as important drug targets for several endocrine disorders [20]. Among them, GCGR and GLP-1R appear to have greater therapeutic potential in diabetes than other members [21–23]. Thus, we focused on GCGR and GLP-1R for further investigation.

GCGR shares high homology with GLP-1R, showing where 54% and 46% sequence identities in the transmembrane and extracellular domains, respectively [24,25]. In addition, the corresponding ligands for GCGR and GLP-1R, glucagon and GLP-1, are also highly conserved in sequence [26]. It has been hypothesized that GLP-1 bound to GCGR and exhibited glucagon-like action in fish, but later it acquired unique incretin functions [27]. In humans, the tissue expression profile of *GCGR* and *GLP-1R*is different. *GCGR* is actively expressed in liver and kidney, whereas *GLP-1R* has relatively high expression in pancreas. This agrees with the fact that glucagon acts primarily on hepatic GCGR to increase plasma glucose, while GLP-1 functions during nutrient ingestion at pancreatic b-cell GLP-1R to enhance insulin synthesis and secretion [25]. These two hormones have significant but opposing roles in regulating glucose homeostasis and are clinically important in the management of diabetes [28]. GLP-1 affects blood glucose, b-cell protection, appetite, and body weight, which has led to the use of multiple GLP-1R agonists for the treatment of type 2 diabetes [29]. In contrast, glucagon is used to treat severe hypoglycemia [30], while GCGR antagonists have been developed to treat type 2 diabetes. Thus, GCGR and GLP-1R show divergent ligand binding profiles and are selective in hormone action, although they are highly homologous and show conserved structures and sequences. Therefore, when GCGR antagonists wrongly target highly homologous GLP-1R in patients with type 2 diabetes, these drugs may lose their efficacy and fail to control the release of glucose by GCGR. Moreover, the unexpected binding of these drugs to GLP-1R might interfere with function of GLP-1R, thus leading to the decreased insulin secretion. As a result, anti-diabetes drugs targeting one of these two paralogous receptors at con- served sites may also target the other one by mistake, resulting in cross-reactivity and generating unexpected side effects.




C

| Alignment position | GCGR | GLP-1R | Property change |
|---|---|---|---|
| 47 | E34 | Q37 | Negatively-charged → Hydrophilic |
| 179 | S150 | A153 | Hydrophilic → Hydrophobic |
| 330 | N291 | D293 | Hydrophilic → Negatively-charged |
| 379 | Q337 | L339 | Hydrophilic → Hydrophobic |
| 387 | F345 | C347 | Hydrophobic → Hydrophilic |
| 432 | F387 | S389 | Hydrophobic → Hydrophilic |
| 450 | K405 | N407 | Positively-charged → Hydrophilic |
| 473 | E427 | Q428 | Negatively-charged → Hydrophilic |

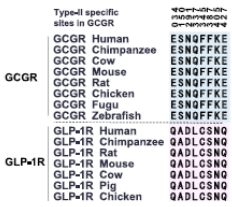



**Figure 1 Analytical pipeline for type-II functional divergence between GCGR and GLP-1R**
**A.** Phylogenetic tree of GCGR and GLP-1R. **B.** Site-specific profile for predicting critical amino acid residues responsible for type-II functional divergence between GCGR and GLP-1R measured by posterior probability $Q_{II}$ (k). **C.** Overview of amino acid changes in the eight predicted sites in type-II functional divergence. **D.** Sequence conservation analysis of the two clusters for GCGR and GLP-1R. GCGR, glucagon receptor; GLP-1R, glucagon-like peptide 1 receptor.

## Usage of type-II functional divergence features as targetable difference of drugs

To avoid undesirable side effects driven by drug interactions with conserved residues of paralogs, we analyzed type-II functional divergence between GCGR and GLP-1R to identify residues conserved in functional constraints but different in physicochemical properties. Based on the neighbor-joining phylogenetic tree (Figure 1A), we estimated the coefficient of type-II functional divergence (denoted by $\theta_{II}$) between GCGR and GLP-1R: $\theta_{II}$=0.236 ± 0.052, which is significantly higher than 0 ($P < 0.001$). A large value of $\theta_{II}$ indicates a high level of type-II functional divergence, and *vice versa*. Rejection of the null hypothesis $\theta_{II}$= 0 means that after gene duplication, the evolutionary rate has become different between the duplicate genes at some residues. Some amino acid residues that were evolutionally conserved in both GCGR and GLP-1R across different species may have radically changed their amino acid properties. Furthermore, we used the posterior probability $Q_{II}(k)$ to identify amino acid residues critical

in type-II functional divergence between these two paralogous genes (Figure 1B). Using an empirical cutoff of $Q_{II}(k) > 0.67$, we identified 8 type-II functional divergence-related residues between paralogous GCGR and GLP-1R. These included E34, S150, N291, Q337, F345, F387, K405, and E427 in GCGR. The site-specific ratio profile indicated that most residues had low posterior ratios and only a small portion of amino acid residues were involved in this type of functional divergence. Moreover, these 8 amino acid residues showed a typical pattern of type-II functional divergence (Figure 1C). They showed a high sequence conservation at paralogous sites (Figure 1D). We sized down the posterior probabilities of these sites and found that using lower posterior probability as cut-off value (such as 0.54) would screen out residues that were not presented in typical conservation pattern of type-II functional divergence (data not shown). Thus, we used these 8 type-II functional divergence-specific sites for further analysis about their roles in paralog discrimination.

### Type-II functionally-divergent residues in binding sites of anti-diabetic drugs

The issues of cross-reactivity arising from paralogs have been long concerned. Identifying paralog-divergent features as tar- getable difference might be helpful in paralog discrimination and has already been implemented in therapeutic drug design [31]. The GCGR antagonist MK-0893 is used to treat patients with type 2 diabetes to substantially reduce fasting and postprandial glucose concentrations [31]. MK-0893 acts at allosteric binding sites of the seven transmembrane helical domain (7TM) in positions among TM5, TM6, and TM7 in GCGR (Figure 2A). TM6 plays a role in splitting the binding sites into two different interaction regions. The TM5-TM6 cleft contains L329, F345, L352, T353, and the alkyl chain of K349, making hydrophobic contacts with one part of MK-0893. On the other hand, the TM6-TM7 section forms polar interactions with the other part of MK-0893 by hydro-gen bonds with K349, S350, L399, N404, and the backbone of K405, as well as additional salt bridge with R346. Thus, the different physicochemical properties function in the binding activity of the dual-nature antagonist MK-0893 to GCGR (Figure 2B). We found that our predicted sites of type-II functional divergence between GCGR and GLP-1R, F345 and K405, were significantly enriched in the binding sites of MK- 0893 to GCGR ($P < 0.05$; chi-square test).

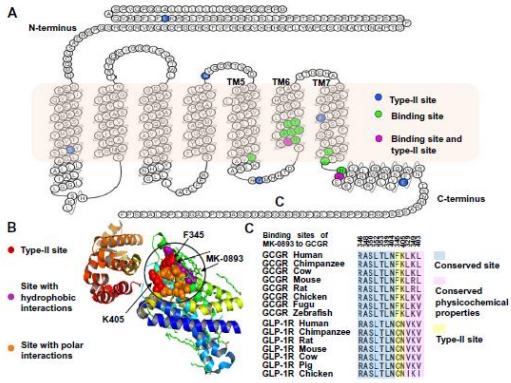



**Figure 2 Paralog-divergent features are considered targetable differences of drugs**
**A.** Snake-plot diagram of GCGR with annotation of important residues. **B.** Different physicochemical properties of bipartite antagonist pocket corresponding to the dual polar/hydrophobic binding cleft in GCGR. **C.** Sequence conservation analysis of 12 binding sites of MK-0893 to GCGR.

To figure out the key difference in paralogous residues between GCGR and GLP-1R, we analyzed the sequence conservation in the binding sites of MK-0893 to GCGR and com- pared them with their equivalent sites in GLP-1R. The results showed that the type-II specific sites F345 and K405 had a radical shift in physicochemical properties, while other binding sites were highly conserved either in functional constraints or physicochemical properties between the two paralogs GCGR and GLP-1R (Figure 2C). F345 and K405, showed a typical pattern of type-II functional divergence. They were both conserved residues in their orthologous gene families, but were different in their physicochemical properties between paralogous GCGR and GLP-1R. F345 was hydrophobic in GCGR but its equivalent site in GLP-1R is hydrophilic. If a molecule of drug is designed to be hydrophobic, it tends to bind to the hydrophobic F345 in GCGR rather than the hydrophilic residue in GLP-1R. Another type-II specific site K405 was positively-charged in GCGR while its equivalent site in GLP-1R was electrically neutral. Thus a molecule of drug designed to be negatively-charged are more likely to interact with positively-charged K405 in GCGR instead of binding to the electrically neutral residue in GLP-1R. Because the physio- chemical properties of amino acids play an important role in the interaction of protein receptors with their ligands (small

molecules, peptides, agonists, and antagonists), changes in their physicochemical nature and conformation may reduce cross-reactivity due to the binding of antagonist drugs to unexpected paralogs. Therefore, determining type-II functional divergence-related sites between two paralogs is effective for identifying targetable differences in therapeutic drug design.

Moreover, we investigated the binding of ligand and agonists to GLP-1R and evaluated the role of type-II functionally-divergent sites between GCGR and GLP-1R in this study. We identified a type-II functional divergence-related residue D293 within human GLP-1R in the second extracellular loop (EC2) (Table S1). D293 showed a typical pattern of type-II functional divergence. This residue is con- served in orthologous gene families of GLP-1R and is functionally important. It had ligand-specific effects on GLP-1 peptide-mediated selective signaling and was critical for agonist-mediated receptor activation [32]. Residue D293 of EC2 directly interacted with key residues in the ligand through hydrogen-bonding interactions (Table S1). A previous study [33] demonstrated that D293A mutation reduced GLP-1 affinity and altered the binding and efficacy of agonists such as oxyntomodulin and exendin-4 [34]. As a type-II functionally important site, D293 in GLP-1R showed different physicochemical properties from its equivalent site N291 in paralogous GCGR. The amino acid property changes from negatively-charged in GLP-1R to electrically neutral in GCGR, which can serve as a selective filter for telling apart GLP-1R from GCGR. Thus, the application of divergence features of type-II functional divergence between these two paralogs is advantageous in this respect.

### Using type-I functional divergence features as targetable difference of drugs

Besides type-II functional divergence, type-I functional divergence between paralogs might also be exploited to achieve targetable differences. We thus investigated the role of residues related to type-I functional divergence in the binding of ligand and agonists to GLP-1R. To do so, we computed the coefficient of type-I functional divergence (denoted by $\theta_I$; $\theta_I = 0$ for the null hypothesis) between GCGR and GLP-1R. We got $\theta_I = 0.4902 \pm 0.1072$, which was significantly higher than 0 ($P<0.001$), indicating the occurrence of type-I functional divergence between two paralogs. We identified a type-I-related residue E294 in the binding sites of GLP-1R. E294 is a functionally important site for the signaling mechanism and receptor activation [32]. It is highly conserved in one cluster of orthologous GLP-1R family but appears as diverse amino acids at paralogous sites in GCGR. Therefore, the type-I functional divergence-related residues might play vital roles in drug binding sites for discrimination of two paralogs for tighter specificity control of drugs.

### Usage of variable residues as targetable difference of drugs

Not all binding sites of drugs have been designed to exploit the type-I or type-II functional divergence features as discriminating factors between paralogs. We therefore investigated more examples to see whether residues other than type-I or type-II functionally-divergent residues can achieve targetable difference between paralogs. We examined GCGR antagonist antibodies mAb1, mAb23, and mAb7 that target the ligand- binding cleft in the N-terminal extracellular domain, where the cleft is typically structurally important in ligand binding for secretin type GPCRs [35]. Our sequence conservation analysis of these antagonists illustrates that most binding-site residues showed significant conservation between paralogous GCGR and GLP-1R ($P = 0.0003$, 0.02, and 0.002 for mAb1, mAb23, and mAb7, respectively; chi-square test. Besides the most conserved residues, there are also some variable residues other than type-II or type-I specific residues in the binding sites. Mutations at these variable residues lead to structural differences such as a shift or changes in orientation of some side chain residues, thus resulting in reduced receptor activation and even prevention of ligand binding [36]. Therefore, these variable residues differ from one paralog to their equivalent sites in another paralog, while other residues in the binding sites are highly conserved either in sequence or in physicochemical properties. This implies that there may be underlying mechanisms involving variable residues in the discrimination of GCGR and GLP-1R.

### Identification of functional divergence of druggable paralogs in GPCRs

Inspired by the usage of functional divergence features in improving drug selectivity between paralog GCGR and GLP-1R, we hypothesized that these features might be applied to other paralogs of GPCRs in drug design. We thus extended to all targetable GPCRs and investigated their types of functional divergence between each paralogous gene pair. We identified 83 drug targets in total in GPCRs superfamily based on the published data on human druggable protein tar- gets (Figure 3). We found that these targets are mainly enriched in rhodopsin, glutamate, and secretin subfamilies, which have been revealed to bind to various types of ligands and are targeted for drug design [17]. Among these 83 targets, 6 and 8 targets belong to the secretin and glutamate subfamilies,

respectively, while others are found in 4 subgroups of rhodopsin subfamily. Interestingly, receptors in adhesion and frizzled/taste2 subfamilies are not found as drug targets. The majority of receptors in these two subfamilies remain orphans, and few attempts have been made to target these two classes.

Based on the two types of functional divergence between each paralogous pair in each subfamily, we found that, within 465 duplicated gene pairs, 267 pairs of paralogs have undergone functional divergence during the evolution. Among them, 67 pairs of paralogs showed only type-I functional divergence and 55 pairs showed only type-II functional divergence, whereas 145 pairs showed both two types of functional divergence (Table S3). Due to the lack of public data on drug binding sites for many targetable receptors in GPCRs family, we were not able to test the functional divergence features of all paralog pairs for verification. However, the site score for probability to be associated with type I or type II functional divergence is shown for each position on the multiple alignment of these paralogous gene pairs (Table S4). We systematically evaluated the large-scale functional divergence of each pair of paralogs in GPCRs to conclude the profiles of type-I or type-II related amino acid residues in every duplicated gene. These observations could be taken into consideration when designing conserved residues as drug binding sites (Table S5).

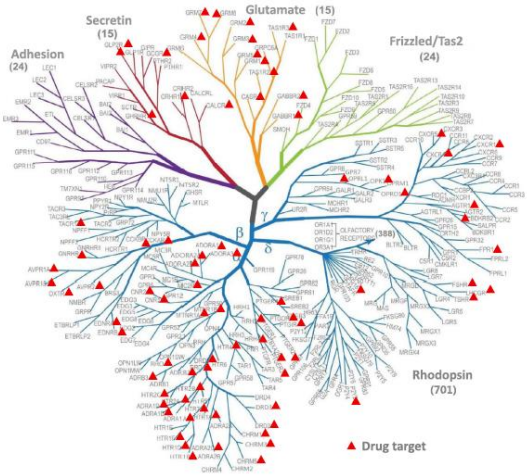



**Figure 3 Identification of drug targets in GPCRs**
Rhodopsin, glutamate, and secretin subfamilies in GPCRs are abundant for drug targets. 83 targetable receptors are plotted on the GPCR tree (courtesy of Vsevolod Katritch and Raymond C Stevens from University of Southern California). GPCR, G-protein coupled receptor.