Statistical Framework for Phylogenomic Analysis of Gene Family Expression Profiles

Xun Gu[1]

Department of Genetics, Development and Cell Biology (GDCB), Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa 50011

Manuscript received July 14, 2003
Accepted for publication January 22, 2004

ABSTRACT

Microarray technology has produced massive expression data that are invaluable for investigating the genome-wide evolutionary pattern of gene expression. To this end, phylogenetic expression analysis is highly desirable. On the basis of the Brownian process, we developed a statistical framework (called the E_0 model), assuming the independent expression of evolution between lineages. Several evolutionary mechanisms are integrated to characterize the pattern of expression diversity after gene duplications, including gradual drift and dramatic shift (punctuated equilibrium). When the phylogeny of a gene family is given, we show that the likelihood function follows a multivariate normal distribution; the variance-covariance matrix is determined by the phylogenetic topology and evolutionary parameters. Maximum-likelihood methods for multiple microarray experiments are developed, and likelihood-ratio tests are designed for testing the evolutionary pattern of gene expression. To reconstruct the evolutionary trace of expression diversity after gene (or genome) duplications, we developed a Bayesian-based method and use the posterior mean as predictors. Potential applications in evolutionary genomics are discussed.

DNA microarray technology can simultaneously monitor the expression levels of thousands of genes across many experimental conditions or treatments (Brown and Botstein 1999), providing us with unique opportunities to investigate the evolutionary pattern of gene regulation (*e.g.*, Wagner 2000; Gu *et al.* 2002; Enard *et al.* 2002; Gu and Gu 2003; Caceres *et al.* 2003; Rifkin *et al.* 2003). To this end, an appropriate statistical framework is highly desirable because current bioinformatic tools for microarray data analysis such as clustering (*e.g.*, Eisen *et al.* 1998) are not sufficient for studying the evolution of gene expression. Besides, the high-level background noise could make many data-driven approaches statistically unreliable (Kerr and Churchill 2001).

A conventional approach to tracing the evolutionary change is to classify each (continuous) expression measure into a binary state (expressed or not expressed; Ferkowicz *et al.* 1998; Force *et al.* 1999; Viale *et al.* 2000; Prince and Pickett 2002; Wang and Gu 2002; Gu and Gu 2003). In spite of its simplicity, the appropriate cut-off for classification depends on data type and experimental design, which may not be obvious in many cases. For microarray data, in particular, assessing statistical significance of gene expression or expression change is still challenging, due to the huge number of genes and very few replicates (Kerr and Churchill 2001; Quackenbush 2001).

Here we focus on the evolution of gene expression after duplications. When the phylogenetic tree of a gene family can be inferred by the sequence data, the pattern of expression profiles among member genes can be modeled as a stochastic process driven by underlying evolutionary mechanisms. This approach has several advantages: (i) Statistical methods such as the likelihood-ratio test can be applied for exploring the evolutionary pattern of gene expression; (ii) evolutionary tracing of expression changes can be predicted by the Bayesian method; (iii) the statistical model can be utilized to study the expression-motif association; and (iv) it may provide a novel approach for phylogenetic reconstruction beyond sequence data.

The goal of this article is to develop a statistical framework for studying the evolution of gene expression after duplications. Since gene expression data can be viewed as continuous characters, the concept of comparative methods for phenotypic evolution is helpful to establish a bridge between massive microarray data and phylogenetic analysis. Starting from a simple random-walk model (Edwards and Cavalli-Sforza 1964; Lynch and Hill 1986), several evolutionary mechanisms are then introduced to characterize the pattern of expression diversity after gene duplication. Since modeling these mechanisms may result in overparameterization, we solve this problem by using the hierarchical modeling approach to derive the likelihood function of gene expressions

[1]*Address for correspondence:* Department of Genetics, Development and Cell Biology, Center for Bioinformatics and Biological Statistics, 332 Science II Hall, Iowa State University, Ames, IA 50011.
E-mail: xgu@iastate.edu

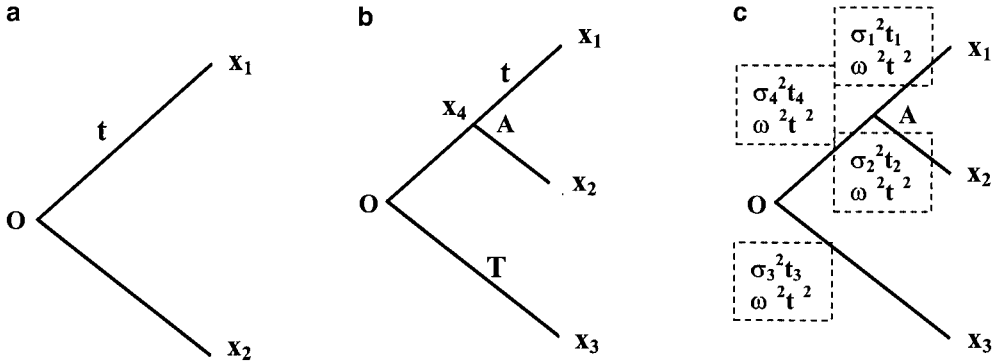

FIGURE 1.—The schematic of a gene family tree. (a) The simplest case of a gene family with two member genes. (b) The three-gene tree for the B model. (c) The three-gene tree for the L and D models. See the text for the interpretation of notations.

when the phylogeny is known. The effect of experimental correlation between microarrays is also considered. Moreover, we propose an "empirical" Bayesian framework to infer the ancestral expression levels at internal nodes, providing a useful tool to reconstruct the evolutionary trace of gene expressions. These newly developed methods are applied for the yeast microarray data.

MODELS AND METHODS

Basic Brownian model (B model): In microarray data, the expression level X of a gene is usually measured by the log-transformed signal intensity, after normalization and bias correction. For the (two-way) cDNA microarray, X measures the relative mRNA abundance to a prespecified condition (control), while for the Affymatrix array, X is a good predictor for the absolute mRNA abundance (QUACKENBUSH 2001).

Given the initial expression level x_0, the Brownian model demonstrates that the expression level $X = x$ after t evolutionary time units follows a normal distribution with mean x_0 and variance $\sigma^2 t$; the density is given by

$$B(x|x_0;\ \sigma^2 t) = \frac{1}{\sqrt{2\pi t\sigma}} e^{-(x-x_0)^2/2\sigma^2 t}. \quad (1)$$

The Brownian model has been used to study the character evolution. Under the mutation-drift model, σ^2 can be interpreted as the mutational variance (LYNCH and HILL 1986; FELSENSTEIN 1988).

Microarray data usually include multiple measurements under various development/tissue stages or experimental treatments. These measurements can be considered as samples from the same stochastic process, *i.e.*, evolutionary replicates (see below), but the initial expression level differs. We assume that the expression level x_0 at the root O (the common ancestor of the gene family) varies among microarray experiments according to a normal distribution $N(x_0;\ \mu,\ \rho^2)$, with the following density:

$$\pi(x_0) = \frac{1}{\sqrt{2\pi\rho}} e^{-(x_0-\mu)^2/2\rho^2}. \quad (2)$$

Gene family with two-member genes: We start from a simple two-member gene family (Figure 1a). Let x_1 and x_2 be the expression levels of two member genes, respectively, and $P(x_1,\ x_2)$ be the joint density. Given the initial value (x_0) of gene expression at the root O, the change of x_1 follows a Brownian process $B(x_1|x_0;\ \sigma^2 t)$, and the change of x_2 follows $B(x_2|x_0;\ \sigma^2 t)$. If the evolution of gene expression is independent between lineages, referred to as the E_0 assumption, we have $P(x_1,\ x_2|x_0) = B(x_1|x_0;\ \sigma^2 t)\, B(x_2|x_0;\ \sigma^2 t)$. From Equations 1 and 2 one can show the joint density of x_1 and x_2 is given by

$$P(x_1,\ x_2) = \int_{-\infty}^{\infty} P(x_1,\ x_2|x_0)\,\pi(x_0)\,dx_0 = N(x_1,\ x_2;\ \boldsymbol{\mu},\ \mathbf{V}); \quad (3)$$

that is, x_1 and x_2 follow a bivariate normal distribution, with the mean vector $\boldsymbol{\mu} = (\mu,\ \mu)$ and the variance-covariance matrix

$$\mathbf{V} = \begin{pmatrix} \rho^2 + \sigma^2 t & \rho^2 \\ \rho^2 & \rho^2 + \sigma^2 t \end{pmatrix}. \quad (4)$$

Under this model, the covariance between x_1 and x_2 equals the ancestral variance (ρ^2) of the root O, indicating that the expression similarity between gene 1 and gene 2 reflects their evolutionary relatedness.

Gene family with three members: For a given three-member gene family (Figure 1b), the joint density of expressions x_1, x_2, and x_3 is derived as follows. Denote the expression level at the ancestral node A by x_4. Let T and t be the evolutionary times of nodes O (the root) and A, respectively. Thus, given the initial value (x_0) at O, the change of x_4 follows $B(x_4|x_0;\ \sigma^2(T-t))$ and the change of x_3 follows $B(x_3|x_0;\ \sigma^2 T)$. Similarly, given the ancestral level x_4, the changes of x_1 and x_2 follow $B(x_1|x_4;\ \sigma^2 t)$ and $B(x_2|x_4;\ \sigma^2 t)$, respectively. According to the Markov property, we obtain the joint density $P(x_1,\ x_2,\ x_3,\ x_4|x_0) = B(x_3|x_0,\ \sigma^2 T)\, B(x_1|x_4;\ \sigma^2 t)\, B(x_2|x_4;\ \sigma^2 t)\, B(x_4|x_0,\ \sigma^2(T-t))$. Since the ancestral expression x_4 is unobservable, we have $P(x_1,\ x_2,\ x_3|x_0) = \int_{-\infty}^{\infty} P(x_1,\ x_2,\ x_3,\ x_4|x_0)\,dx_4$. After some math, one can show that the joint density

$$P(x_1,\ x_2,\ x_3) = \int_{-\infty}^{\infty} P(x_1,\ x_2,\ x_3|x_0)\,\pi(x_0)\,dx_0 \quad (5)$$

is a three-variate normal density $N(x_1, x_2, x_3; \boldsymbol{\mu}, \mathbf{V}_B)$, with the mean vector $\boldsymbol{\mu} = (\mu, \mu, \mu)'$, and the variance-covariance matrix

$$\mathbf{V}_B = \begin{pmatrix} \rho^2 + \sigma^2 T & \rho^2 + \sigma^2(T - t) & \rho^2 \\ \rho^2 + \sigma^2(T - t) & \rho^2 + \sigma^2 T & \rho^2 \\ \rho^2 & \rho^2 & \rho^2 + \sigma^2 T \end{pmatrix}. \quad (6)$$

The proof can be outlined as follows. Under the Brownian model, $P(x_1, x_2, x_3|x_0)$ is obviously normally distributed. Hence the normality of $P(x_1, x_2, x_3)$ is guaranteed when $\pi(x_0)$ is a normal density. Using the technique of conditional expectations, one can derive $\boldsymbol{\mu}$ and $\mathbf{V}_B$ easily.

Lineage-specific expression diversity (L model): The B model assumes that the expression divergence of a gene family is driven mainly by small and additive genetic drifts (random effects), with a constant rate measured by σ^2, or by mutational variance under the drift-mutation model (Lynch and Hill 1986). In this sense, the B model can be considered as the "neutral-evolution" model of gene expression or the "expression clock."

A more general model is the L model, under which the evolutionary rate σ^2 of expression diversity is lineage specific. Let $\sigma_k^2 t_k$ be the variance component in the kth branch ($k = 1, \ldots, 4$) in the three-gene phylogeny (Figure 1c) and t_k be the corresponding evolutionary time. Similar to the B model, we have shown that the joint density of x_1, x_2, and x_3,

$$P(x_1, x_2, x_3) = \int_{-\infty}^{\infty} \pi(x_0) B(x_3|x_0, \sigma_3^2 t_3) \times \left[\int_{-\infty}^{\infty} B(x_1|x_4, \sigma_1^2 t_1) B(x_2|x_4, \sigma_2^2 t_2) B(x_4|x_0, \sigma_4^2 t_4)\, dx_4 \right] dx_0, \quad (7)$$

is a three-variate normal density $N(x_1, x_2, x_3; \boldsymbol{\mu}, \mathbf{V}_L)$, with the mean vector $\boldsymbol{\mu} = (\mu, \mu, \mu)'$, and the variance-covariance matrix

$$\mathbf{V}_L = \begin{pmatrix} \rho^2 + \sigma_1^2 t_1 + \sigma_4^2 t_4 & \rho^2 + \sigma_4^2 t_4 & \rho^2 \\ \rho^2 + \sigma_4^2 t_4 & \rho^2 + \sigma_2^2 t_2 + \sigma_4^2 t_4 & \rho^2 \\ \rho^2 & \rho^2 & \rho^2 + \sigma_3^2 t_3 \end{pmatrix}. \quad (8)$$

Obviously, when $\sigma_k^2 = \sigma^2$ for all $k = 1, \ldots, 4$, the L model is reduced to the B model.

Directional trend of expression diversity (D model): The lineage-specific (L) Brownian model can be further extended to the Brownian model with directional trend (the D model). That is, given the initial value x_0 at $t = 0$, the stochastic process of gene expression change after t time units is described as $B(x|x_0; \lambda t, \sigma^2 t)$, *i.e.*, a normal distribution with mean $x_0 + \lambda t$ and variance $\sigma^2 t$; λ is called the trend coefficient or the coefficient of directional selection (Felsenstein 1988). Apparently, the L model is the special case of $\lambda = 0$. Since the trend coefficient (λ) of a gene varies among microarray samples, we treat λ as a random variable that follows a normal distribution $f(\lambda) = N(\lambda; \bar{\lambda}, \omega^2)$. The "marginal" Brownian motion is therefore given by

$$B(x|x_0; \bar{\lambda} t, \sigma^2 t + \omega^2 t^2) = \int_{-\infty}^{\infty} B(x|x_0; \lambda t, \sigma^2 t) f(\lambda)\, d\lambda = N(x; x_0 + \bar{\lambda} t, \sigma^2 t + \omega^2 t^2); \quad (9)$$

i.e., it is a normal distribution with mean $x_0 + \bar{\lambda} t$ and variance $\sigma^2 t + \omega^2 t^2$. Note that the variance under the D model is a quadratic function of time t.

The joint density of gene expressions under the D model can be derived by adding lineage-specific directional trends to the L model. For a three-member gene family, Equation 7 can be modified as

$$P(x_1, x_2, x_3) = \int_{-\infty}^{\infty} \pi(x_0) B(x_3|x_0; \bar{\lambda}_3 t_3, \sigma_3^2 t_3 + \omega_3^2 t_3^2) \times \left[\int_{-\infty}^{\infty} B(x_1|x_4; \bar{\lambda}_1 t_1, \sigma_1^2 t_1 + \omega_1^2 t_1^2) B(x_2|x_0; \bar{\lambda}_2 t_2, \sigma_2^2 t_2 + \omega_2^2 t_2^2) \times B(x_4|x_0; \bar{\lambda}_4 t_4, \sigma_4^2 t_4 + \omega_4^2 t_4^2)\, dx_4 \right] dx_0, \quad (10)$$

where $\bar{\lambda}_k$ and ω_k^2 ($k = 1, \ldots, 4$) are the mean and variance of directional trends in the kth branch, respectively (Figure 1c). Similar to the derivation of the L model, one can show that $P(x_1, x_2, x_3)$ is a three-variate normal density $N(x_1, x_2, x_3; \boldsymbol{\mu}, \mathbf{V})$, with the mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)'$, where $\mu_1 = \mu + \bar{\lambda}_1 t_1 + \bar{\lambda}_4 t_4$, $\mu_2 = \mu + \bar{\lambda}_2 t_2 + \bar{\lambda}_4 t_4$, and $\mu_3 = \mu + \bar{\lambda}_3 t_3$. The variance-covariance matrix $\mathbf{V}_D = \mathbf{V}_L + \mathbf{W}_D$, where $\mathbf{V}_L$ is given by Equation 8, and the matrix of directional trend $\mathbf{W}_D$ is given by

$$\mathbf{W}_D = \begin{pmatrix} \omega_1^2 t_1^2 + \omega_4^2 t_4^2 & \omega_4^2 t_4^2 & 0 \\ \omega_4^2 t_4^2 & \omega_2^2 t_2^2 + \omega_4^2 t_4^2 & 0 \\ 0 & 0 & \omega_3^2 t_3^2 \end{pmatrix}. \quad (11)$$

Dramatic shift of expression diversity (S model): The D model we developed above assumes that the change of gene expression is continuous with time t (the gradual evolution). However, a dramatic shift (positively or negatively) in gene expression may happen shortly after gene duplication and then remain little changed, *i.e.*, the process of "punctuated equilibrium" (Hansen and Martins 1996). The S model assumes that after gene duplication the expression level has shifted z and z' units in two lineages, respectively (Figure 2). It is further assumed that two shift variables (z and z'), which are independent of each other, vary among microarray experiments according to normal distributions $N(z; \delta, s^2)$ and $N(z'; \delta', s'^2)$, respectively.

For a three-gene family, the joint density of x_1, x_2, and x_3 under the S model is derived as follows. After the first gene duplication, the initial expression level (x_0) at the root O is immediately shifted to $x_0 + z_O$ in one lineage

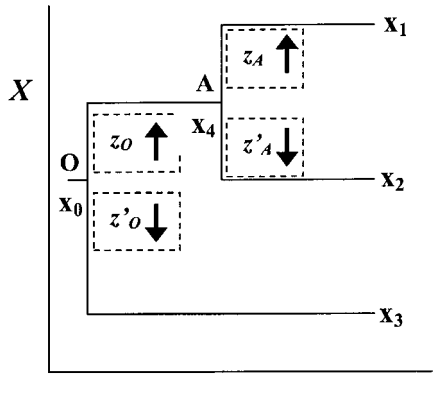

FIGURE 2.—Schematic of the dramatic shift S model, which shows a version of punctuated equilibrium. See the text for the interpretation of notations.

and to $x_0 + z'_O$ in another lineage (Figure 2). Hence, the change of x_4 follows a Brownian process $B(x_4|x_0 + z_O, \sigma_4^2 t_4)$, and the change of x_3 follows $B(x_3|x_0 + z'_O, \sigma_3^2 t_3)$. After the second gene duplication, the expression level (x_4) at node A is immediately shifted to $x_4 + z_A$ in one lineage and to $x_4 + z'_A$ in another lineage (Figure 2). In the same manner, we have $B(x_1|x_4 + z_A, \sigma_1^2 t_1)$ for the change of x_1 and $B(x_2|x_4 + z'_A, \sigma_2^2 t_2)$ for the change of x_2. For simplicity, we assume no gradual drift in each branch. Then, the joint density conditional of x_0, z_O, z'_O, z_A, z'_A can be written as follows:

$$\begin{aligned} P(x_1, x_2, x_3|x_0, z_O, z'_O, z_A, z'_A) &= B(x_3|x_0 + z'_O, \sigma_3^2 t_3) \\ &\times \int_{-\infty}^{\infty} B(x_1|x_4 + z_A, \sigma_1^2 t_1) \\ &\times B(x_2|x_4 + z'_A, \sigma_2^2 t_2) \\ &\times B(x_4|x_0 + z_O, \sigma_4^2 t_4)\, dx_4. \end{aligned} \tag{12}$$

Apparently, the L model (Equation 7) is a special case when $z_A = z'_A = 0$ and $z_O = z'_O = 0$. The next step is to integrate out all shift variables z_A, z'_A, z_O, and z'_O. Letting $\phi(z)$ be the normal density for any shift variable z, we have

$$P(x_1, x_2, x_3|x_0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(x_1, x_2, x_3|x_0, z_O, z'_O, z_A, z'_A)\phi(z_A)\phi(z'_A)\,dz_A dz'_A\right] \times \phi(z_O)\phi(z'_O)\,dz_O dz'_O. \tag{13}$$

And finally, the initial expression at the root (x_0) is integrated out according to Equation 2; *i.e.*,

$$P(x_1, x_2, x_3) = \int_{-\infty}^{\infty} \pi(x_0) P(x_1, x_2, x_3|x_0)\, dx_0. \tag{14}$$

We have shown that $P(x_1, x_2, x_3)$ follows a multivariate normal distribution. The mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)'$ is given by $\mu_1 = \mu + \delta_A + \delta_O$, $\mu_2 = \mu + \delta'_A + \delta_O$, and $\mu_3 = \mu + \delta'_O$; and the variance-covariance matrix $\mathbf{V}_S = \mathbf{V}_L + \mathbf{W}_S$, where $\mathbf{V}_L$ is given by Equation 8 and the shift matrix $\mathbf{W}_S$ is given by

$$\mathbf{W}_S = \begin{pmatrix} s_A^2 + s_O^2 & s_O^2 & 0 \\ s_O^2 & s_A'^2 + s_O'^2 & 0 \\ 0 & 0 & s_O'^2 \end{pmatrix}. \tag{15}$$

It has been shown that if the gradual drift (D) model is considered, one can show that the variance-covariance matrix is given by $\mathbf{V}_S = \mathbf{V}_L + \mathbf{W}_D + \mathbf{W}_S$, where $\mathbf{W}_D$ is given by Equation 11.

The general likelihood function under the E_0 model: The joint density for a three-member gene family can be extended to any n-member gene family when a rooted phylogeny is given. We have shown that the joint density of $\mathbf{x} = (x_1, \ldots, x_n)$ follows a multivariate normal distribution, $N(x_1, \ldots, x_n; \boldsymbol{\mu}, \mathbf{V})$; with the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix $\mathbf{V}$, the density is given by

$$P(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n |\mathbf{V}|^{1/2}} \exp\left\{-\frac{(\mathbf{x} - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})}{2}\right\}. \tag{16}$$

This result can be proved by using the principle of mathematical induction: Given that Equation 16 holds for $n = 3$, we have shown that if Equation 16 holds for $n = k$, it must be true for $n = k + 1$. A complete proof under any phylogenetic tree is given by Z. ZHANG and X. GU (unpublished results). We call it the E_0 model because it assumes independent evolution between lineages. The main results are summarized as follows.

The B model: The mean vector $\boldsymbol{\mu} = (\mu, \ldots, \mu)'$, and the variance-covariance matrix $\mathbf{V}_B$ is given by $\mathbf{V}_B = \rho^2\mathbf{J} + \sigma^2 T\mathbf{U}$, where any ijth element of matrix $\mathbf{J}$ is 1, and T is the age of the common ancestor (the root O). The tree-topology-related matrix $\mathbf{U}$ is defined as $U_{ii} = 1$ and $U_{ij} = 1 - a_{ij}$ $(i \neq j)$, where $a_{ij} = t_{ij}/T$; t_{ij} is the age of the common ancestor of genes i and j. More explicitly, the ijth element of $\mathbf{V}_B$ is given by

$$V_{ij} = \begin{cases} \rho^2 + \sigma^2 T, & \text{if } i = j \\ \rho^2 + \sigma^2 T(1 - a_{ij}) & \text{if } i \neq j. \end{cases} \tag{17}$$

The L model: Since the L model allows specific σ_k^2 for each branch k, one cannot separate the topology matrix $\mathbf{U}$ from other components. Instead, the ijth element of $\mathbf{V}_L$ is

$$V_{ij} = \begin{cases} \rho^2 + \sum\limits_{k \in x_i} \sigma_k^2 t_k, & \text{if } i = j \\ \rho^2 + \sum\limits_{k \in (x_i, x_j)} \sigma_k^2 t_k & \text{if } i \neq j, \end{cases} \tag{18}$$

where t_k is the evolutionary time of branch k, the subscript notation $k \in x_i$ runs over all branches in the lineage from the root O to gene x_i, and $k \in (x_i, x_j)$ runs over all branches shared by x_i and x_j since the root O (Figure 3).

The D model: It has been shown that μ_i, the mean of each x_i, is equal to μ plus the sum of mean drifts in branches from the root O to gene x_i; *i.e.*,

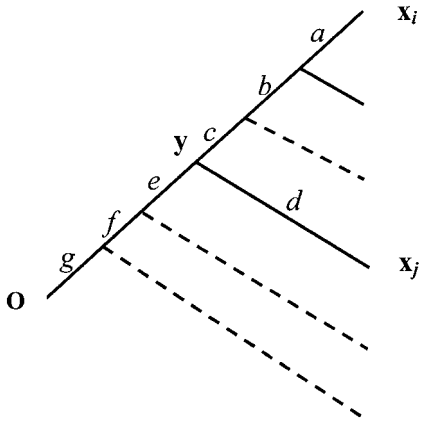

FIGURE 3.—Schematic of a gene family tree, where y is the expression level at node A, the common ancestor of genes i and j. Thus, $k \in x_i = (a, b, c, e, f, g)$ for all branches from the root O to gene i; $k \in (x_i, x_j) = (e, f, g)$ for all branches from the root O to A; and also $k \in (x_i, y) = (e, f, g)$ for all branches from the root O to A.

$$\mu_i = \mu + \sum_{k \in x_i} \bar{\lambda}_k t_k. \tag{19}$$

The variance-covariance matrix can be written as $\mathbf{V}_D = \mathbf{V}_L + \mathbf{W}_D$, where $\mathbf{V}_L$ is given by Equation 18. The ijth element of matrix $\mathbf{W}_D$ is given by $W_{ii} = \Sigma_{k \in x_i} \omega_k^2 t_k^2$ and $W_{ij} = \Sigma_{k \in (x_i, x_j)} \omega_k^2 t_k^2$ $(i \neq j)$ (Figure 3). Therefore, $\mathbf{V}_D$ can be expressed as

$$V_{ij} = \begin{cases} \rho^2 + \sum_{k \in x_i} \sigma_k^2 t_k + \sum_{k \in x_i} \omega_k^2 t_k^2 & \text{if } i = j \\ \rho^2 + \sum_{k \in (x_i, x_j)} \sigma_k^2 t_k + \sum_{k \in (x_i, x_j)} \omega_k^2 t_k^2 & \text{if } i \neq j, \end{cases} \tag{20}$$

which is a quadratic function of evolutionary time of each branch.

The S model: Under the general S model, the mean of each x_i is the sum of the mean gradual drifts (over branches) and the mean shifts (over the ancestral nodes) from root O to x_i; that is,

$$\mu_i = \mu + \sum_{k \in x_i} \bar{\lambda}_k t_k + \sum_{m \in x_i} \delta_m, \tag{21}$$

where $m \in x_i$ runs over all ancestral nodes between the root O and x_i (root O included but x_i not included). The variance-covariance matrix $\mathbf{V}_S$ can be expressed as

$$V_{ij} = \begin{cases} \rho^2 + \sum_{k \in x_i} \sigma_k^2 t_k + \sum_{k \in x_i} \omega_k^2 t_k^2 + \sum_{m \in x_i} S_m^2 & \text{if } i = j \\ \rho^2 + \sum_{k \in (x_i, x_j)} \sigma_k^2 t_k + \sum_{k \in (x_i, x_j)} \omega_k^2 t_k^2 + \sum_{k \in (x_i, x_j)} S_m^2 & \text{if } i \neq j, \end{cases} \tag{22}$$

where for each ancestral node m, $S_m = s_m$ or s_m' for two follow-up branches, respectively. Let A_{ij} be the most recent common ancestral node of x_i and x_j. Thus, the notation $m \in (x_i, x_j)$ runs over all ancestral nodes between the root O and the node A_{ij} (root O included but A_{ij} not included).

Ancestral gene expression inference: Bayesian approach: Ancestral state reconstruction along a phylogenetic tree is at the center of comparative methods in evolutionary biology, for both morphological and molecular characters (HARVEY and PAGEL 1991; YANG *et al.* 1995; SCHLUTER *et al.* 1997; GOLDING and DEAN 1998). The massive microarray data make it possible to reconstruct an ancestral expression pattern that is useful to trace the evolutionary changes of gene regulation.

Single-node ancestral inference: This method provides a fast Bayesian procedure to infer ancestral expression profile because each time it deals with one ancestral node and then runs over the tree. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be the observed expression pattern and y be the expression level at the ancestral node of interest (Figure 2). According to Bayes' rule, the posterior density $P(y|x_1, \ldots, x_n)$ is computed as follows:

$$P(y|x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n, y)}{P(x_1, \ldots, x_n)}. \tag{23}$$

From Equation 16, we know $P(x_1, \ldots, x_n)$ is an n-variate normal density. To derive the numerator of Equation 23, we use a three-gene family for illustration, where $y = x_4$. Note that $P(x_1, x_2, x_3, x_4|x_0) = B(x_3|x_0)B(x_1|x_4)B(x_2|x_4)B(x_4)$, according to the Markov property. Since $P(x_1, x_2, x_3, x_4) = \int_\infty^\infty B(x_3|x_0)B(x_1|x_4)B(x_2|x_4)B(x_4)\pi(x_0)\,dx_0$, similar to the derivation of $P(x_1, x_2, x_3)$, we show that $P(x_1, x_2, x_3, x_4)$ is a 4-variate normal density.

In the general case, let $M = n + 1$ and regard the ancestral level y as an additional variable x_{n+1}. It has been shown that $P(x_1, \ldots, x_n, y)$ is an $(n + 1)$-variate normal density, denoted by $N(x_1, \ldots, x_n, y; \boldsymbol{\mu}, \mathbf{V_M})$. The extended variance-covariance matrix $\mathbf{V_M}$ has the following structure: If $1 \leq i, j \leq n$, the ijth element of $\mathbf{V_M}$ is equal to that of $\mathbf{V}$, *e.g.*, Equation 22 under the general S model. For any i, $n + 1$th element, $i = 1, \ldots, n + 1$, it is given by

$$V_{i,n+1} = \begin{cases} \rho^2 + \sum_{k \in y} \sigma_k^2 t_k + \sum_{k \in y} \omega_k^2 t_k^2 & \text{if } i = n + 1 \\ \rho^2 + \sum_{k \in (x_i, y)} \sigma_k^2 t_k + \sum_{k \in (x_i, y)} \omega_k^2 t_k^2 & \text{if } i \neq n + 1 \end{cases} \tag{24}$$

and $V_{n+1,i} = V_{i,n+1}$, where the subscript notation $k \in y$ runs over all branches in the lineage from the root O to the ancestral node y, and $k \in (x_i, y)$ runs over all branches shared by x_i and y since the root O (Figure 3). For simplicity, we assume $\boldsymbol{\mu} = (\mu, \ldots, \mu)'$.

Hence, it becomes obvious that the posterior density $P(y|x_1, \ldots, x_n)$ is a normal density. Let $\mathbf{C} = \mathbf{V}_M^{-1}$; c_{ij} is the ijth element of $\mathbf{C}$. After some algebra we obtain

$$P(y|x_1, \ldots, x_n) = \frac{1}{\sqrt{2\pi}\sigma_{y|x}} \exp\left\{ -\frac{1}{2\sigma_{y|x}^2} \left[y - \mu + \sum_{i=1}^{n} \frac{c_{i,n+1}}{c_{n+1,n+1}} (x_i - \mu) \right]^2 \right\}, \tag{25}$$

where $\sigma_{y|x}^2 = 1/c_{n+1,n+1}$ is the (posterior) variance of y. That is, the posterior mean of y conditional of $\mathbf{x} = (x_1, \ldots, x_n)'$ is given by

$$E[y|x_1, \ldots, x_n] = \beta_0 + \sum_{i=1}^{n} \beta_i x_i, \quad (26)$$

where $\beta_i = -c_{i,n+1}/c_{n+1,n+1}$ and $\beta_0 = \mu(1 + \Sigma_{i=1}^{n}\beta_i)$. Apparently, the posterior mean prediction for the ancestral gene expression is a linear function of current gene expressions.

Joint ancestral inference: To explore the joint evolutionary pattern of expression changes after gene duplications, the single-node method may not be sufficient. Therefore we develop an approach for joint ancestral expression inference. For a gene family with n member genes, there are m ancestral nodes when the phylogenetic tree is given. Let $\mathbf{x} = (x_1, \ldots, x_n)'$ and $\mathbf{y} = (y_1, \ldots, y_m)'$ be the vectors of current and ancestral expression levels, respectively, and $M = n + m$. The (extended) $M \times M$ variance-covariance matrix for $(\mathbf{y}', \mathbf{x}')$ is denoted by $\mathbf{V}_M$. We have shown that $P(\mathbf{y}, \mathbf{x})$ is an M-dimensional multinormal density. It follows that the joint posterior density of ancestral nodes $\mathbf{y}$,

$$P(\mathbf{y}|\mathbf{x}) = \frac{P(\mathbf{y}, \mathbf{x})}{P(\mathbf{x})} = \frac{N(\mathbf{y}, \mathbf{x}; \boldsymbol{\mu}, \mathbf{V}_M)}{N(\mathbf{x}; \boldsymbol{\mu}, \mathbf{V})}, \quad (27)$$

is also $m \times m$ multinormal, that is, $P(\mathbf{y}|\mathbf{x}) = N(\mathbf{y}; \boldsymbol{\mu}_{y|x}, \Sigma_{y|x})$, where $\boldsymbol{\mu}_{y|x} = (\mu_{y_1|x}, \ldots, \mu_{y_m|x})'$ is the posterior mean vector of the ancestral nodes, and $\Sigma_{y|x}$ is the $m \times m$ posterior variance-covariance matrix of $y_1, \ldots, y_m$.

To obtain useful analytical results for numerical calculation, we partition the matrix $\mathbf{V}_M$ as

$$\mathbf{V}_M = \begin{bmatrix} \mathbf{A} & \mathbf{H} \\ \mathbf{H}' & \mathbf{V} \end{bmatrix}, \quad (28)$$

where $\mathbf{H}$ and $\mathbf{A}$ are $m \times n$ and $m \times m$ matrices, respectively. The matrix $\mathbf{H}$ is the ancestral-current expression covariances and $\mathbf{A}$ is the variance-covariance matrix among ancestral nodes. Thus, the inverse of the matrix $\mathbf{V}_M$ can be written as

$$\boldsymbol{\Lambda}_M = \begin{bmatrix} \mathbf{A} & \mathbf{H} \\ \mathbf{H}' & \mathbf{V} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{yy} & \boldsymbol{\Lambda}_{yx} \\ \boldsymbol{\Lambda}'_{yx} & \boldsymbol{\Lambda}_{xx} \end{bmatrix}, \quad (29)$$

where $\boldsymbol{\Lambda}_{xx}$, $\boldsymbol{\Lambda}_{xy}$, and $\boldsymbol{\Lambda}_{yy}$ are $n \times n$, $m \times n$, and $m \times m$ matrices, respectively. It has been shown that

$$\begin{aligned} \textstyle\sum_{y|x} &= \boldsymbol{\Lambda}_{yy}^{-1} \\ \boldsymbol{\mu}_{y|x} &= \boldsymbol{\mu} - \boldsymbol{\Lambda}'_{yx}\boldsymbol{\Lambda}_{yy}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \end{aligned} \quad (30)$$

(X. Gu and Z. Zhang, unpublished results).

Implementation and data analysis: *Multiple microarray experiments:* The microarray data collection for evolutionary analysis can be outlined as follows: (1) Multiple (N) microarray data sets are downloaded from the Stanford microarray database and (2) a relational database is established to extract the expression profiles of any given gene family. A typical data set for a three-gene family is

$$\begin{bmatrix} \text{gene family} \\ \text{gene A} \\ \text{gene B} \\ \text{gene C} \end{bmatrix} = \begin{bmatrix} k = 1 & k = 2 & \ldots & k = N \\ 1.9 & 0.2 & \ldots & -0.2 \\ -1.4 & 0.9 & \ldots & 1.2 \\ 0.7 & 1.5 & \ldots & 2.0 \end{bmatrix},$$

where the kth column represents the expression profile across the gene family in the kth microarray experiment, and the ith row represents the expression profile of gene i across the N microarray experiments. For our interest, we view multiple microarrays as *evolutionary replicates*, on the basis of the following considerations.

For a given gene family, the whole set of *cis*- and *trans*-regulatory elements is diversified following gene duplications. Ideally, these r regulatory elements can be represented as a binary string denoted by $\mathbf{h} = h_1, \ldots, h_r$. Each duplicate gene has a unique representation of $\mathbf{h}$. These "aligned" regulatory elements, similar to aligned nucleotide sites, can be viewed as *the evolutionary replicates* from a stochastic process. Microarray samples of cells/tissues at various developmental stages or under experimental treatments show characteristic expression profiles that reflect the differences in $\mathbf{h}$ among member genes. Thus, the expression profile of gene i at the kth microarray data can be conceptually written as $x_{i,k} = \psi_k(\mathbf{h})$, although ψ is yet little known. In this regard, multiple microarray experiments are also *evolutionary replicates* but the sampling property needs to be addressed carefully.

Likelihood under the i.i.d. assumption: The likelihood function can be easily built under the i.i.d. assumption that microarray experiments are independently, identically distributed, which has been used in many previous evolutionary studies (*e.g.*, Wagner 2000; Gu *et al.* 2002). For N microarray experiments, let $\mathbf{x}_k = (x_{1,k}, \ldots, x_{n,k})$ be the expression pattern of an n-member gene family at the kth experiment. When the phylogeny is given, the likelihood for gene expressions can be written as

$$L(\mathbf{V}, \boldsymbol{\mu}|\text{data}) = \prod_{k=1}^{N} P(\mathbf{x}_k; \boldsymbol{\mu}, \mathbf{V}). \quad (31)$$

The maximum-likelihood estimates of the parameters can be obtained by the Newton-Raphson iteration method, and the sampling variance of each estimate is approximated by the inverse of the information matrix.

Expression branch lengths: The number of unknown parameters under the E_0 model could be larger than the degrees of freedom, which makes the model statistically infeasible. One useful solution is to define the expression branch length E_k for any branch k along the phylogeny, that is,

$$E_k = \sigma_k^2 t_k + \omega_k^2 t_k^2 + S_k^2, \quad (32)$$

where the subscript k under S_k^2 is for the initial node of the kth branch. Then, the variance-covariance matrix under the general E_0 model (Equations 16 and 22) can be simplified as

$$V_{ij} = \begin{cases} \rho^2 + \sum_{k \in x_i} E_k & \text{if } i = j \\ \rho^2 + \sum_{k \in (x_i x_j)} E_k & \text{if } i \neq j. \end{cases} \quad (33)$$

Hence, the statistical framework with parameters (ρ^2, E_k, $k = 1, 2, \ldots$) becomes a general model for constructing useful likelihood-ratio tests. We have implemented our methods using the statistical software package S-plus.

Likelihood under experimental correlations: Hundreds of microarray experiments, say, for the yeast, include time-course experiments, tissue/developmental stages, cell cycles, stress-response experiments, different environmental conditions, mutants (gene deletions), etc. (Eisen *et al.* 1998). The overall expression profiles with similar types of conditions or treatments are more similar to each other; *e.g.*, two adjacent sampling points in a time-course assay are usually highly correlated. Because of these experimental correlations, the i.i.d. assumption seems to be unrealistic. That is, the sample of expression profiles of a gene family is not only phylogenetically but also experimentally dependent.

As the first-order approximation, we model the experimental correlation of the microarray data as the overall correlations among microarray experiments. Let **D** be the $N \times N$ matrix of experimental correlations; that is, the diagonal element is 1, while the off-diagonal element is the coefficient of correlation between any two microarray experiments. For the microarray chip that includes in total C genes (*e.g.*, $C = 5500$ for the yeast genome), **D** can be estimated by the standard approach over all C genes. Then one can show that the expression profiles of the gene family $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$ follow a multivariate normal distribution, with the large $[n \times N] \times [n \times N]$ variance-covariance matrix $\mathbf{V} \otimes \mathbf{D}$. The likelihood function is therefore given by

$$L(\mathbf{V} \otimes \mathbf{D}, \boldsymbol{\mu} | \text{data}) = P(\mathbf{X}; \boldsymbol{\mu}, \mathbf{V} \otimes \mathbf{D}). \quad (34)$$

Apparently, Equation 34 is reduced to Equation 31 when $\mathbf{D} = \mathbf{I}$. When the dimension $n \times N$ is not large, a numerical algorithm based on the matrix decomposition is developed. Here we show a simple example.

Wolfe and Shields (1997) have identified 452 duplicate genes that were from the yeast genome duplication. Using Chu *et al.*'s (1998) cDNA microarray data, seven-point time-course experiments during the sporulation, we calculated the experimental correlation matrix **D** (Chu *et al.* 1998). Obviously, two adjacent time points usually show a high correlation, while it is low between distant points. The variance and covariance elements of **V** for a duplicate pair are given by $V_{11} = \rho^2 + E_1$, $V_{22} = \rho^2 + E_2$, and $V_{12} = \rho^2$. These unknown parameters (ρ^2, E_1, and E_2) are then estimated using Equations 31 and 34, respectively. We use the estimated coefficient of correlation $R_{12} = \hat{\rho}^2 / \sqrt{(\hat{\rho}^2 + \hat{E}_1)(\hat{\rho}^2 + \hat{E}_2)}$ to compare their performance. All 452 pairs are plotted in Figure 4, which shows that for the dense-sampled time course, the effect of experimental correlation is nontrivial.

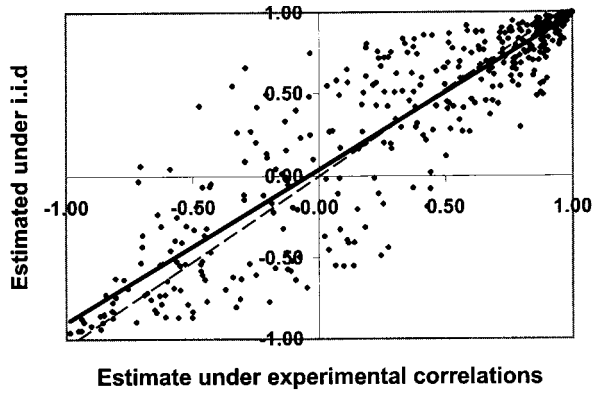

Figure 4.—Expression correlation during yeast sporulation between genome duplication pairs. ML estimates under the i.i.d. assumption or under the experimental correlation are shown.

Yeast glutamyl- and glutaminyl-tRNA synthetases gene family: An example: The glutamyl- and glutaminyl-tRNA synthetases (GlnS) family has three member genes (YGL245w, YOR168w, and YOL033w). Phylogenetic analysis has shown that YGL245w and YOR168w are more closely related (Figure 5). The cell-cycle yeast microarray data are used (Eisen *et al.* 1998).

Under the general E_0 model, the likelihood under the i.i.d. assumption results in the maximum-likelihood (ML) estimates of $\hat{\mu} = -0.058 \pm 0.031$, $\hat{\rho}^2 = 0.053 \pm 0.014$, $\hat{E}_1 = 0.100 \pm 0.023$, $\hat{E}_2 = 0.062 \pm 0.020$, $\hat{E}_3 = 0.099 \pm 0.031$, and $\hat{E}_4 = 0.079 \pm 0.020$. The maximum log-likelihood value is -146.19. Next we consider the likelihood function considering the experimental correlations. We first compute the matrix **D** of microarray experiments. Using the ML estimates under i.i.d. as initial values, we obtain $\hat{\mu} = -0.050$, $\hat{\rho}^2 = 0.055$, $\hat{E}_1 = 0.112$, $\hat{E}_2 = 0.068$, $\hat{E}_3 = 0.104$, and $\hat{E}_4 = 0.061$. It seems that the likelihood under the i.i.d. assumption is useful for fast and large-scale analysis.

Using the molecular clock approach, we have approximately dated the relative time of first gene duplication (between YGL245w/YOR168w and YOL033w) as 2.2 (to the *Escherichia coli*/yeast split time), and the second one as 1.27. Thus, under the B model, we have obtained the ML estimates (under i.i.d.) of $\hat{\mu} = -0.084 \pm 0.033$, $\hat{\rho}^2 = 0.057 \pm 0.015$, and $\hat{\sigma}^2 = 0.047 \pm 0.005$. The maximum log-likelihood value under the B model is -154.13. Apparently, the likelihood-ratio test shows that the B model, or the "expression clock," is rejected at the significance level of 0.001; $\chi^2_{[3]} = 2(154.13 - 146.19) = 15.88$.

As shown in Figure 6, the expression profile of the ancestral ancestor of YGL245w-YOR168w has been inferred by the Bayesian method. Therefore, one can infer lineage-specific changes after gene duplication (the derived characters) from the ancestral expression pattern.

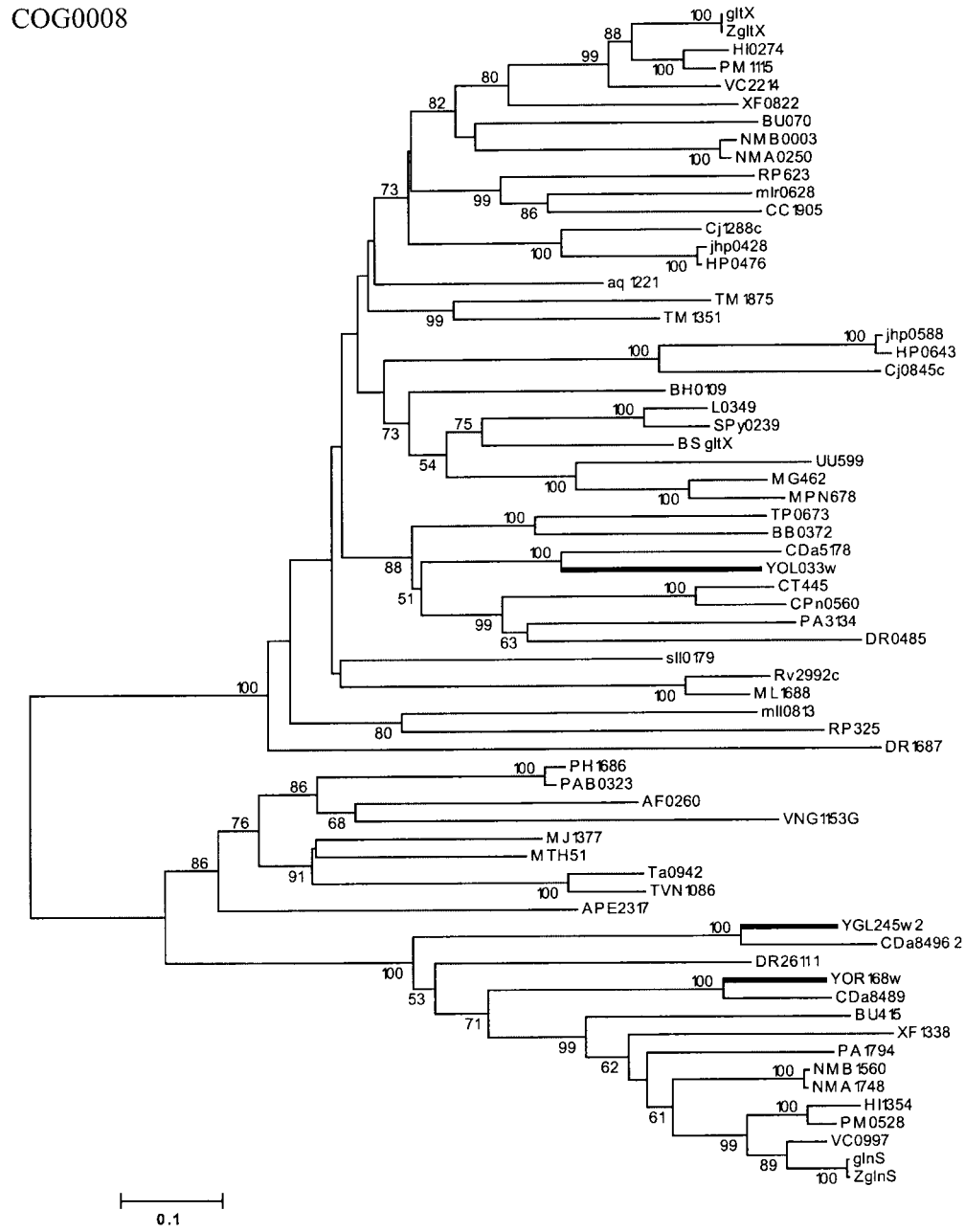

FIGURE 5.—The phylogenetic tree of the GlnS gene family, inferred from the multialignment of amino acid sequences including eukaryotes and prokaryotes. The neighbor-joining method was used.

For instance, changes of expression level in each lineage after gene duplication will be very informative when other genome data (*e.g.*, putative regulatory motifs) are available.

Yeast gene families: We have conducted a large data analysis for yeast gene families to understand the evolution of expression after gene duplications. Amino acid sequences of yeast gene families, as well as homologous genes from nearly 30 complete genomes, were downloaded from the COG database (http://www4.ncbi.nlm.nih.gov/COG/). The phylogenetic tree of each gene family is inferred by the neighbor-joining method (SAITOU and NEI 1987). Furthermore, the duplication time relative to yeast/*E. coli* split can be (approximately) estimated by using the molecular clock approach under the inferred tree, similar to GU *et al.* (2002). In total 276 yeast microarray data are collected for each gene family; the likelihood of Equation 31 is used for the analysis.

For a three-member gene family when the root can be reliably determined, the likelihood-ratio test (LRT) is used to test the expression clock hypothesis (Figure 7). The null hypothesis is $E_1 = E_2$. The log of likelihood ratio approximately follows a χ^2 with d.f. = 1 so that one can determine the significance level. We conducted the LRT for 60 three-member yeast gene families; for them the relative duplication times estimated from sequence data are treated as known. Overall, 42 (70%) gene families showed that the null hypothesis is rejected

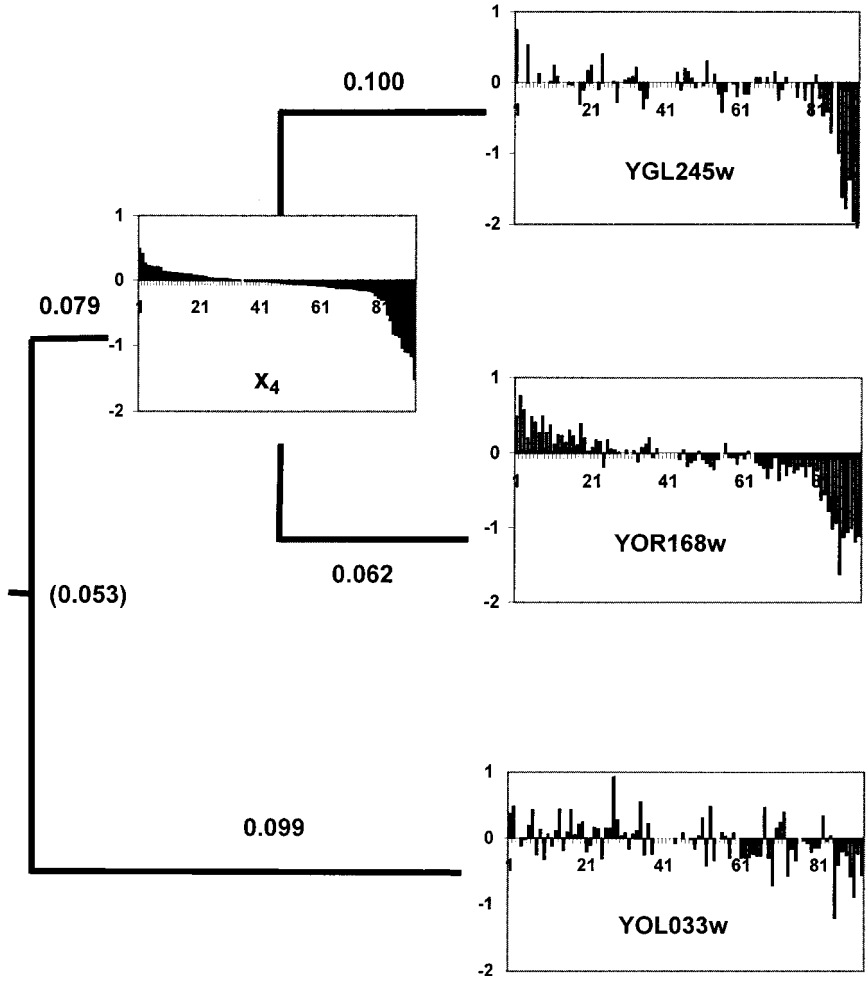

FIGURE 6.—Phylogenetic expression analysis of the yeast GlnS family. The expression branch length (b_k^x) is indicated in each branch. The inferred ancestral expression profile for the common ancestor of YGL245W and YOR165W is presented.

at the 0.05 significance level and 39 (65%) gene families at the 0.01 significance level. Therefore, one may conclude that, after gene duplication, an unequal rate for expression divergence in each duplicate gene is a common pattern. Examples in Figure 8 show more dramatic expression changes than sequence substitutions after gene duplications.

DISCUSSION

In this article we have developed a statistical framework to explore the expression divergence during the gene family evolution. Several data analyses have shown the potential for studying many interesting problems in evolutionary genomics, *e.g.*, the evolution of gene expression specificity, the evolutionary fate of duplicate genes, ancestral expression inference, and the coevolution between expression and regulatory motif or coding sequence. Our method is also useful for understanding the population-quantitative genetic basis of expression evolution after gene duplication.

Comparative methods: Statistical methods for the analysis of continuous morphological data have been studied for decades to deal with phylogenetic-dependent sampling (HARVEY and PAGEL 1991; MADDISON 1994). These methods can be roughly classified into phylogenetically independent contrasts (PIC; FELSENSTEIN 1988), the

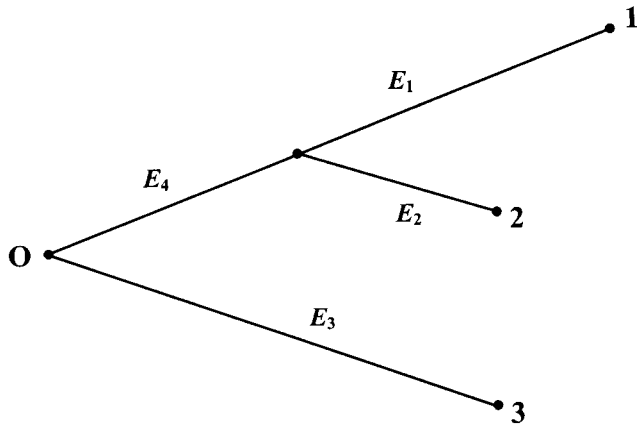

FIGURE 7.—Likelihood-ratio tests (LRT) for unequal expression divergence between gene 1 and gene 2; gene 3 is used as an outgroup.

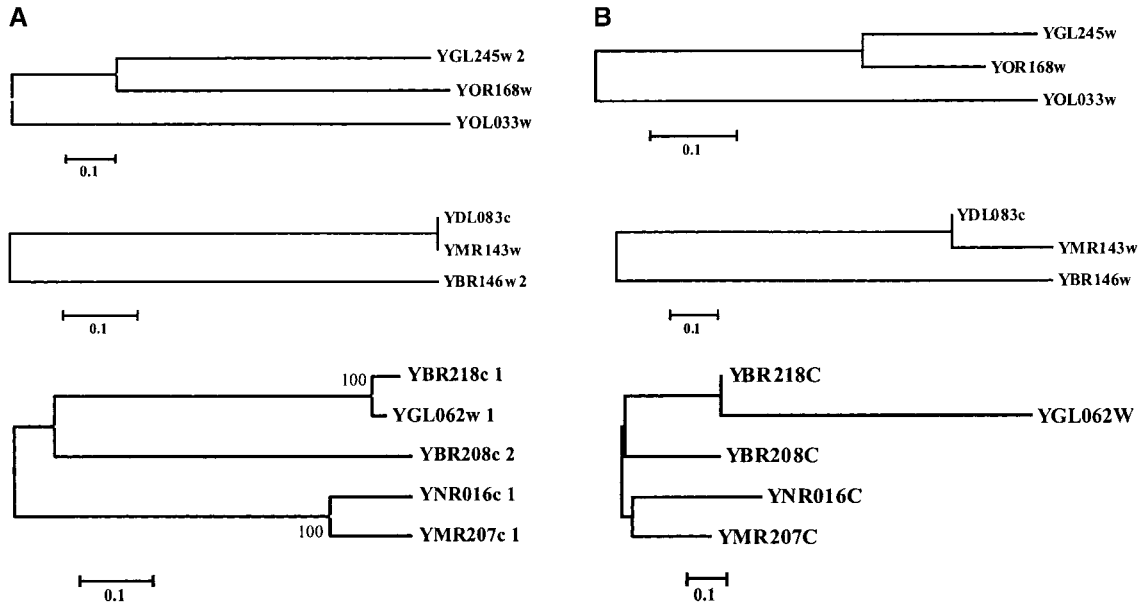

FIGURE 8.—Example of unequal expression divergence after gene duplications. The three gene families are COG008, COG0103, and COG0508. For each gene family, the left side (A) is inferred from amino acid sequences and the branch lengths in the right side (B) represent expression branch length.

phylogenetic autocorrelation (PA) method (CHEVERUD *et al.* 1985), and phylogenetic generalized least-squares (PGLS; GRAFEN 1989), as well as many variants; see ROHLF (2001) for a recent detailed review and critiques. Here we briefly discuss the connection between the newly developed method and these existing methods.

We have noted that all these methods share a similar theoretical ground. In fact, PIC, PA, and PGLS methods require the known "expected variance-covariance matrix" **V** for incorporating the phylogenetic structure. Except for some early studies, the popular assumption is $\mathbf{V} = \sigma^2\mathbf{C}$, where the matrix **C** is determined by the phylogeny and timescales. Essentially, it is equivalent to the B model. SCHLUTER *et al.* (1997) introduced a simple prior for σ^2 to relax the unrealistic assumption. There has been long-term controversy about which one should be used and exactly how it can be applied (PAGEL 1993; ROHLF 2001). ROHLF (2001) did not consider any complex form of **V**, because it is more arbitrary in implementation. The central theme of our study is to model the expected variance-covariance matrix **V** for exploring the evolutionary pattern of expression divergence, which can be estimated under appropriate statistical procedures. Since the likelihood-ratio test clearly shows a strong lineage-specific mode of expression evolution, the conventional PIC, PA, and PGLS methods that assume a constant Brownian motion may be oversimplified for microarray data analysis.

Inference of the ancestral expression pattern provides an efficient approach to reconstruct the evolutionary trace of expression diversity after gene (or genome) duplications. The so-called maximum-likelihood method for inferring the ancestral state of continuous characters is to minimize the (weighted) sum of differences over all branches (*e.g.*, SCHLUTER *et al.* 1997). For the three-gene tree (Figure 1b), it is given by

$$Q = [(x_1 - x_A)^2/t + (x_2 - x_A)^2/t + (x_A - x_O)^2/(T - t) + (x_3 - x_O)^2/T]/\sigma^2. \quad (35)$$

Obviously, the estimation requires known evolutionary times t and T. Since the assumption of the constant Brownian model does not hold in general for the microarray data, the ancestral inference in Equation 35 could be highly biased. Our empirical Bayesian approach provides in general more accurate inference for the ancestral expression pattern because the lineage-specific effect is well taken into consideration. Finally, we point out that the Bayesian method of SCHLUTER *et al.* (1997) is to assume uniform priors for the mean vector and the log of σ^2, respectively, which did not specifically address the lineage-specific effect.

Models extended for expression data: In the future, we will study how to implement the methodology of PIC, PA, or PGLS for the microarray data. For instance, the PGLS regression model for gene family expression may improve the efficiency of the expression-motif search (*e.g.*, BUSSEMAKER *et al.* 2001; BLANCHETTE and TOMPA 2002). This is because the regulatory processing of a gene family can be traced back to a single ancestral regulatory modular so that the heterogeneity problem can be partially avoided. For a given microarray experiment, the expression profile of a gene family $\mathbf{x} = (x_1, \ldots, x_n)'$ can be written as $\mathbf{x} = \mathbf{Hb} + \boldsymbol{\varepsilon}$, where the matrix **H** represents explanatory variables, *e.g.*, the *cis*-regulatory motif structure (predicted or known, num-

ber of copies, locations, etc.), and some *trans*-elements that may influence the expression level. The variance-covariance matrix of the error term $\boldsymbol{\varepsilon}$, denoted by $\boldsymbol{\Sigma}$, is given by

$$\sum = \mathbf{V} - \rho^2\mathbf{I}, \tag{36}$$

where $\mathbf{V}$ is given by Equation 32; ρ^2 should be removed since the root state is treated as a fixed value when a single data set is considered.

Similarly, the PA method can be written as $\mathbf{x} = \alpha\mathbf{S}\mathbf{x} + \boldsymbol{\varepsilon}'$, where α is the autocorrelation coefficient, $\mathbf{S}$ is an $n \times n$ (tree-dependent) connection matrix (each row of which sums up to one), and $\boldsymbol{\varepsilon}'$ is error terms (Cheverud *et al.* 1985). Instead of the arbitrary construction of $\mathbf{S}$, we suggest estimating $\mathbf{S}$ from $\boldsymbol{\Sigma}$ in Equation 36 with (row) normalization. Moreover, for a gene family, it would be very interesting to detect any *functional correlation* between two microarray experiments. We may modify the PIC method (Felsenstein 1988) to correct not only phylogenetic dependence but also experimental correlation as well as the lineage-specific effect.

Experimental errors and others: We have recognized that the current statistical framework involves several assumptions that need to be examined carefully. The first one is the normal assumption. By default, we assume that fold change of gene expression follows a Brownian-based process during evolution. Further investigation is needed to test the robustness of normal assumption when other data normalization procedures are adopted (Quackenbush 2001). Second, we should consider the sources of experimental errors. Let σ_e^2 be the variance of experimental errors. Then the variance-covariance matrix $\mathbf{V}$ of Equation 33 should be modified as

$$V_{ij} = \begin{cases} \rho^2 + \sum_{k \in x_i} E_k + \sigma_e^2 & \text{if } i = j \\ \rho^2 + \sum_{k \in (x_i, x_j)} E_k & \text{if } i \neq j. \end{cases} \tag{37}$$

We study the performance of our method when the "background noise" is strong (Kerr and Churchill 2001). For instance, the LRT tests may be liberal so a more stringent significance level is suggested. Since the variance of experimental errors does not depend on the tree topology, it can be estimated when replicates for a single microarray experiment become the standard protocol.

The author thanks Zhongqi Zhang for assisting in numerical analysis. This work was supported by a National Institutes of Health grant.

LITERATURE CITED

Blanchette, M., and M. Tompa, 2002 Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res. **12:** 739–748.

Brown, P. O., and D. Botstein, 1999 Exploring the new world of the genome with DNA microarrays. Nat. Genet. **21** (Suppl.): 33–37.

Bussemaker, H. J., H. Li and E. D. Siggia, 2001 Regulatory element detection using correlation with expression. Nat. Genet. **27:** 167–174.

Caceres, M., J. Lachuer, M. A. Zapala, J. C. Redmond, L. Kudo *et al.*, 2003 Elevated gene expression levels distinguish human from non-human primate brains. Proc. Natl. Acad. Sci. USA **100:** 13030–13035.

Cheverud, J. M., M. M. Dow and W. Leutengger, 1985 The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. Evolution **39:** 1335–1351.

Chu, S., J. DeRisi, M. Eisen, J. Mulholland, D. Botstein *et al.*, 1998 The transcriptional program of sporulation in budding yeast. Science **282:** 699–705.

Edwards, A. F. W., and L. L. Cavalli-Sforza, 1964 Reconstruction of evolutionary trees, pp. 67–76 in *Phenetic and Phylogenetic Classification*, No. 6, edited by W. H. Heywood and J. McNeill. Syst. Assoc., London.

Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein, 1998 Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA **95:** 14863–14868.

Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig *et al.*, 2002 Intra- and interspecific variation in primate gene expression patterns. Science **296:** 340–343.

Felsenstein, J., 1988 Phylogenies and quantitative characters. Annu. Rev. Ecol. Syst. **19:** 445–471.

Ferkowicz, M. J., M. C. Stander and R. A. Raff, 1998 Phylogenetic relationships and developmental expression of three sea urchin Wnt genes. Mol. Biol. Evol. **15:** 809–819.

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

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

Grafen, A., 1989 The phylogenetic regression. Philos. Trans. R. Soc. London B **326:** 119–157.

Gu, J., and X. Gu, 2002 Induced gene expression in human brain after the split from chimpanzee. Trends Genet. **19:** 63–65.

Gu, X., Y. Wang and J. Gu, 2002 Age-distribution of human gene families showing equal roles of large and small-scale duplications in vertebrate evolution. Nat. Genet. **31:** 205–209.

Gu, Z., D. Nicolae, H. Lu and W.-H. Li, 2002 Rapid divergence in expression between duplicate genes inferred from microarray data. Trends Genet. **12:** 609–613.

Hansen, T. F., and E. P. Martins, 1996 Translating between microevolutionary process and macroevolutionary patterns: the correlation structure of interspecific data. Evolution **50:** 1404–1417.

Harvey, P., and M. D. Pagel, 1991 *The Comparative Method in Evolutionary Biology.* Oxford University Press, Oxford.

Kerr, M. K., and G. A. Churchill, 2001 Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc. Natl. Acad. Sci. USA **98:** 8961–8965.

Lynch, M., and W. G. Hill, 1986 Phenotypic evolution by neutral mutation. Evolution **40:** 915–935.

Maddison, D. R., 1994 Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters. Annu. Rev. Entomol. **39:** 267–292.

Pagel, M., 1993 Seeking the evolutionary regression coefficient: an analysis of what comparative methods measure. J. Theor. Biol. **164:** 191–205.

Prince, V. E., and B. F. Pickett, 2002 Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. **3:** 827–837.

Quackenbush, J., 2001 Computational analysis of microarray data. Nat. Rev. Genet. **2:** 418–427.

Rifkin, S. A., J. Kim and K. P. White, 2003 Evolution of gene expression in the *Drosophila melanogaster* subgroup. Nat. Genet. **33:** 138–144.

Rohlf, F. J., 2001 Comparative methods for the analysis of continuous variables: geometric interpretations. Evolution **55:** 2143–2160.

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

Schluter, D., T. Price, A. Mooers and D. Ludwig, 1997 Likelihood of ancestor states in adaptive radiation. Evolution **51:** 1699–1711.

Viale, A., A. Courseaux, F. Presse, C. Ortola, C. Breton *et al.*, 2000 Structure and expression of the variant melanin-concentrating hormone genes: only PMCHL1 is transcribed in the developing human brain and encodes a putative protein. Mol. Biol. Evol. **17:** 1626–1640.

Wagner, A., 2000 Decoupled evolution of coding region and mRNA expression patterns after gene duplication: implications for the neutralist-selectionist debate. Proc. Natl. Acad. Sci. USA **97:** 6579–6584.

Wang, Y., and X. Gu, 2002 Evolutionary analysis for developmental profile of gene family expression in central nervous system. Inf. Sci. **145:** 215–223.

Wolfe, K., and D. C. Shields, 1997 Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387:** 708–713.

Yang, Z., S. Kumar and M. Nei, 1995 A new method of inferences of ancestral mucleotide and amino acid sequences. Genetics **141:** 1641–1650.

Communicating editor: Y.-X. Fu