2004

Statistical analysis of gene expression profiles

Zhongqi Zhang
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd
Part of the Statistics and Probability Commons

Recommended Citation
https://lib.dr.iastate.edu/rtd/1135

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Statistical analysis of gene expression profiles

by

Zhongqi Zhang

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Co-majors: Statistics; Bioinformatics and Computational Biology

Program of Study Committee:
Kenneth J. Koehler, Co-major Professor
Xun Gu, Co-major Professor
Alicia L. Carriquiry
Drena Dobbs
Karin Dorman
Dan Nettleton

Iowa State University
Ames, Iowa
2004

Copyright © Zhongqi Zhang, 2004. All rights reserved.
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
This is to certify that the doctoral dissertation of

Zhongqi Zhang

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

Co-major Professor

Signature was redacted for privacy.

For the Co-major Program

Signature was redacted for privacy.

For the Co-major Program
# TABLE OF CONTENTS

## CHAPTER I: GENERAL INTRODUCTION 1
- Introduction 1
- The Brownian-Based Model for Gene Expression Divergence during Genome Evolution 8
- Dissertation Organization 20
- References 21

## CHAPTER II: A FAST METHOD FOR INFERRING ANCESTRAL EXPRESSION STATES AFTER GENE DUPLICATIONS 25
- Abstract 25
- Introduction 25
- Ancestral Expression Inference 27
- Fast Algorithm for Ancestral Expression Inference 29
- Implementation 32
- An Example: Ancestral Expression Inference for Yeast Enolase Gene Family 33
- Discussion 38
- References 40
- Appendix 44

## CHAPTER III: RECONSTRUCTING THE PHYLOGENETIC TREES FROM GENE EXPRESSION DATA USING THE DISTANCE METHODS 47
- Abstract 47
- Introduction 47
- Methods 50
- Enolase (Eno) Gene Family of *Saccharomyces cerevisiae*: an Example 56
- Discussion 62
- References 65
- Appendix 68
CHAPTER IV: HOW MUCH EXPRESSION DIVERGENCE AFTER YEAST GENE DUPLICATION COULD BE EXPLAINED BY REGULATORY MOTIF EVOLUTION?

Abstract 71
Introduction 71
Acknowledgement 79
Box 1. Computational analysis for yeast genome datasets 80
References 85

CHAPTER V: GENERAL CONCLUSIONS 89

General Discussion 89
Future Research 90

ACKNOWLEDGEMENTS 92
CHAPTER I: GENERAL INTRODUCTION

Introduction

Current research in gene expression analysis
As the new century began, biological science advanced into the genomic era. The complete genome sequences of several major organisms have been finished. At the same time, large-scale gene expression analysis allows us to simultaneously monitor the gene expression profiles of tens of thousands of genes. Both sequencing techniques and microarray techniques have been improved over the last decade and resulted in the accumulation of large amounts of data. How to cope with such massive amounts of data? How to obtain and interpret the information conveyed by those analyses? All those questions are waiting for solutions from biologists, as well as statisticians and computer scientists.

At the current stage, most gene expression data analyses can be divided into two major classes: gene clustering and gene expression comparison. For biologists, finding genes of same or similar function is both exciting and time-consuming. Since direct measure of function is currently not easy, scientists are turning to gene expression measurement as an indicator of gene activity. Therefore, a group of genes with similar expression profiles are usually thought to function similarly or at least to act in related pathways. In gene expression clustering, the most common and straightforward method is hierarchy clustering (Eisen et al. 1998; Wen et al. 1998), usually based on the expression correlation between gene expression profiles, where gene pairs with the highest expression similarity are grouping together to form clusters, then cluster and cluster are condensed to form a new cluster until all genes are classified as major groups or “super clusters”. Other popular clustering techniques include singular value decomposition (SVD; Fellenberg et al. 2001; Alter 2000), support vector machines (SVM; Brown et al. 2000) and Bayes clustering (Ramoni M.F. et al. 2002). Clustering itself can be used to identify the function of genes by checking the function annotation of other genes with known function in the same cluster. Moreover, maybe more importantly, clustering provides a group of gene with similar function/expression which can be used as the input dataset for other analysis, such as motif identification analysis. The
underlying assumption here is that genes with similar expression profiles share the same response to gene regulation, for example, have the same binding sites for the transcription factors (TF). The second major research area in expression analysis is to identify genes which are differentially expressed under different experimental conditions. Compared with the expression profile under the normal conditions, genes with significant expression change under certain condition indicate that their function is linked with that condition. In other words, finding a gene with an expression change is likely to reveal its function roles. Some statistical methods used in detecting differentially expressed genes include the standard t-test (Pan 2002), regression modeling (Thomas 2001), mixture modeling (Tusher 2001), etc. For those methods, it is required that the microarray data have replicate measurements, which is not always fulfilled in practice due to the current high cost of microarray chips.

**Sequence divergence, expression divergence and function divergence**

For most biological studies, the ultimate goal is to understand the function of genes, how they interact with each other and how they response to environment changes. However, finding the explicit gene function is not an easy task, at least at the current research stage, due to the complicated nature of gene network and metabolism pathways. Therefore, many scientists turn to some easy-to-handle objects which can shed light on gene functions. In the central dogma of molecular biology, DNA is first transcribed to mRNA and then mRNA is translated into protein. As the final products, proteins are usually the executors of biological activities. Unlike 3D structures, amino acid or its corresponding DNA sequences are relatively easy to handle because of their simple composition and one dimensional form. For years, molecular evolutionists have been looking for functional divergence through the study of sequence divergence within gene families. It is believed that only a few amino acid residue changes are responsible for the function divergence after gene duplication (Golding and Dean 1998). Therefore, by comparing the sequence similarity among genes within the family, it is possible to identify amino acid residues which are crucial to the function divergence (Henikoff 1997; Bork and Koonin 1998). For instance, Gu proposed a “two-state” model for detecting function divergence by looking for significant evolutionary rate shifts after gene
duplication or speciation (Gu 1999). This method has successfully identified some residues which are important for the proper function of the proteins.

If sequence divergence can be used to infer function divergence, why not also use expression divergence? In general, gene expression is an indicator of gene activity; and the expression level under certain conditions is usually associated with its function role under those conditions. For example, during yeast sporulation, sets of genes are induced sequentially corresponding to different sporulation stages – early, middle, mid-late, and late (Chu et al. 1998), and some of them have been experimentally confirmed to possess functions related to different stages of sporulation. To explore expression divergence, we proposed an approach to infer the ancestral gene expression profile and compared the expression changes along the progeny lineages after gene duplication. Statistical testing is adopted to distinguish “true” divergences and those differences simply attributable to stochastic noise. Moreover, such expression divergence can be mapped to specific experiment conditions, revealing potential functional roles of the duplicate genes.

**Gene duplication theory and phylogenetic studies**

Gene duplication is believed to play a significant role in providing vital source for novel functions (Ohno 1970; Hughes 1994). It has been shown that not only can one gene be duplicated into two genes (tandem duplications), but one chromosome or even the whole genome (polyploidization) can be doubled as well. After gene duplication event, one gene becomes two identical/similar copies. Because of the existence of gene redundancy, the selection pressure is temporarily released for either one gene or both. Such release in selection pressure enables the duplicate(s) to go through some level of ‘free’ mutations, therefore, providing a way for the new function to evolve. During gene duplication, the original function of the ancestral gene should be preserved in at least one copy, otherwise, even the potentially beneficial mutation can be deleterious to the organism if such mutation is likely to destroy the original function of the gene.

Although the concept of gene duplication has been accepted by people for a rather long time, the theory underlying the gene duplication event has long been under debate. One reason is that, little sequence information was available at that time, especially genome
sequences. (Another obvious reason is that, we still do not have the time machine.) Based on limited data, the classic model proposes that, for the two duplicates to be preserved, one copy must undergo some beneficial mutation, i.e. gaining some new functions and therefore avoid degeneration caused by function redundancy. Such theory is referred as “Neofunctionalization”. As a complement to neofunctionalization, nonfunctionalization assumes that, after gene duplication, one of the two duplicates will go through some level of degenerated mutations and become a pseudogene. In the classical model, there are two fates for the duplicates, gaining new function or being silenced. Many studies have demonstrated that degenerated mutation is always dominant over beneficial mutation, i.e. most mutation events tend to reduce the original gene function or total destroy the original function, and it is very rare that a mutation can increase the survival rate of the organism. Therefore, it was postulated that the number of functional duplicate genes should be very limited within the genome.

As science advances into the twenty-first century, the genome sequence projects become one of the most fast-growing areas in biological science. More and more organisms’ entire genome sequences have been completed in the past decade. After genome sequence analysis, an unexpected proportion of the genes within all sequenced genomes has been identified as duplicate genes. For example (Prince 2002), more than 15% of human genes are duplicated genes; in zebrafish the percentage is at least 20%; and in Arabidopsis and maize the percentage of duplicated genes are as high as 35%. Therefore, the classical model of neofunctionalization is challenged since it fails to explain the existence of high portion of functional duplicate genes in the genome. In 1999, a new model, called subfunctionalization or duplication-degeneration-complementation (DDC) model, is proposed by Force et al. (Force 1999; Lynch and Force 2000), which aim to explain what classical model fails to explain. In the subfunctionalization model, degenerated mutation is no longer always a “gene killer”; instead, it could be a “gene preserver” as well. The scientific basis for subfunctionalization is the multifunctional nature of individual genes, where one gene can have more than one functional module. Here, functional module refers to different functional domains of the gene product or different regulatory elements in different tissues/conditions. Figure 1 shows a diagram of subfunctionalization, where functional modules are exemplified
using different regulatory elements corresponding to different gene expression patterns. After gene duplication event, null mutations happen on different, however complementary, regulatory regions in both duplicates. Since the regulation of the gene expression is complementary between both copies, the two duplicates together are required to recapitulate the original expression pattern (function) of the ancestral gene, preventing the gene pair from non-functionalization due to function redundancy.

Figure 1. The diagram of subfunctionalization gene duplication model. Empty symbols refer to null mutations and small block arrows indicate sites where complementary null mutations occur.

Aside from duplication theories, the study of evolution also highly relies on the tree-making techniques, or phylogenetic inference method. Application of tree-making methods in biology started with the comparative analysis at the morphological level, where the evolutionary relationship among species or organisms is reconstructed based on their morphological characteristics. After DNA sequencing technology becomes accessible, DNA/protein sequences were used to infer their evolutionary relationships, called molecular phylogenies. Because of the simple composition of the sequence molecules, molecular

---

1 Such complementary degeneration is very rare in general.
phylogenetic inference has advanced dramatically during the past several decades. Among several commonly used methods are distance methods (Sokal and Michener 1958; Cavalli-Sforza and Edwards 1967; Saitou and Nei 1987), maximum parsimony methods (Henning 1966; Eck and Dayhoff 1966; Hartigan 1973) and maximum likelihood methods (Felsenstein 1981; Kishino et al. 1990). (See introduction section in chapter III for more details.) Most tree-making methods are for reconstructing a bifurcating tree, where one ancient gene is duplicated into two progeny genes after every duplication event.

Note that, although gene expression analysis is one of the most fast-developing research areas in science, few studies are focused on its applications in the context of evolution, i.e. expression divergence as discussed in the previous section, which leads us to present some thoughts in chapter II and III.

**DNA motif identification**

![DNA motif identification diagram](image)

Figure 2. The binding of transcription factor (oval) to the DNA motif (short rectangle) induces the gene (long rectangle) transcription.

One important aspect in biological research is to understand the mechanism of gene regulation, where transcription factors (TF) bind specifically to some short sequences, called (DNA) motifs, in the upstream region of the DNA coding sequences, and such interaction between the TF and the motifs results in either enhancing or repressing the gene expression...
activity (figure 2), i.e. the transcription of mRNA from genomic DNA. Starting in the 1990s, after a large amount of DNA sequences were available, identification of motifs became an important and exciting research field in biological sciences, as well as in computer sciences and statistics. In the late 1990’s, several complete genome sequences were finished and large scale gene expression measurement – microarray – became available, motif identification then became even more promising because it successfully combines the sequence information with the expression information. First, DNA sequences are clustered into groups based on their gene expression similarity, then within each co-expressed group, a motif identification algorithm is applied to find the over-represented motifs in the upstream regions based on the assumption that co-expressed genes share similar regulatory motif structures.

So far, there are many methods available in motif identification. Based on their underlying strategies, most of these methods can be divided into two major classes: “word counting” methods (also called “enumerative” or “exhaustive” methods) and “probabilistic” sequence models (also called “alignment” methods; Thijs 2001; Ohler 2001). Word counting methods (Helden 1998; Helden 2000; Jensen 2000) are based on the frequency of oligonucleotides in the upstream regions. The over-represented motifs are then identified from the background noise by comparing the observed frequency to the expected frequency. In the probabilistic methods (Lawrence 1990 & 1993; Bailey 1993 & 1994), a motif model is proposed, usually based on the position probability matrix, and the model parameters are estimated using the maximum likelihood (ML). Based on which method is used in the maximum likelihood estimation, the probabilistic methods can be further classified as several sub-methods, such as Expectation Maximization (EM) and Gibbs sampling methods.

In general, the motif structures in the promoter regions of the coding sequences, in conjunction with the transcription factors, will affect the expression activity of genes. However, the interaction between the transcription factors and the motifs should not be considered as the only factor determining the expression process (Wray et al. 2003). Indeed, many researches have shown that the effect of transcription factors in expression regulation is usually context dependent (Lemon and Tjian 2000; Courey 2001) which mostly involves the interaction among transcription factors or between transcription factors and their cofactors, as well as the chromatin structures. For most TF, the presence or absence of
cofactors is crucial for their proper function (Chen and Courey 2000). It is observed that some motifs can be recognized by different TF, some motif sequences are overlapping in the promoter region or close enough to affect the binding of one motif to another. In those situations, the gene expression is largely determined by the competition among different motifs. For most motif-finding algorithms, the underlying assumption is that there is a pattern common to all the sequences in the input dataset. In practice, such input dataset is usually collected based on the expression similarity among genes. As already noticed, the gene expression profiles may or may not be consistently linked to the motif structure; therefore, the input dataset may be “contaminated” by the presence of upstream sequences that do not contain the common motifs.

The Brownian-Based Model for Gene Expression Divergence during Genome Evolution

Most of my research about statistical analysis of gene expression data is based on the Brownian-Based model developed by Gu (2004). In the following paragraphs, I will review some of the details about this model, including the implementation of this model to real dataset followed by some discussions.

The General Brownian-Based Model

Figure 3: 3-Gene Phylogeny with Expression Divergence
For an n-gene rooted phylogeny, the observed genes are labeled through 1 to n, with corresponding expression values $x_1$ to $x_n$. The most ancient ancestor is annotated as root 0 with expression $x_0$, while all other intermediate ancestral (internal) genes are referred as gene $n+1, ..., 2n-2$. For example, figure 3 is a 3-gene phylogeny, with root 0 and three external (observed) genes: 1, 2 and 3. Gene 4 is the interior (unobserved) node that is also the direct/nearest ancestor of genes 1 and 2. The branch between a gene (either interior or exterior) and its nearest ancestor is labeled with that gene’s number. As in figure 3, the branch between gene 1 and 4 is branch 1 and branch between gene 4 and 0 is branch 4. $t_i$ is used to refer the evolution time associated with branch $i$, as shown in figure 3.

Within the n-gene phylogeny, the gene expression ($x_i$) given the expression of its nearest ancestor ($x_k$) is normally distributed with conditional mean $x_k$ and its conditional variance is modeled as a Brownian process ($\sigma^2 t_i$). For two descendant genes after gene duplication, it is also assumed that their gene expressions (given their ancestral expression) are independent. A $z_i$ component was included in the conditional mean, representing the sum of all kinds of expression changes after gene duplication, as shown below,

$$ P(x_i|x_k, z_i; \sigma^2 t_i) = \frac{1}{\sqrt{2\pi t_i \sigma_i}} e^{-\frac{(x_i-x_k-z_i)^2}{2\sigma_i^2 t_i}}. $$

It also assumes the expression at root ($x_0$) has a prior distribution of

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

Next, $z_i$ and $z_j$, the two gene shifts after the same gene duplication events, are assumed to follow a bivariate normal distribution with density

$$ \phi(z_i, z_j|\delta_i, \delta_j, \sigma_i^2, \sigma_j^2, r_{ij}) = \frac{1}{2\pi \sqrt{(1-r_{ij}^2)\sigma_i \sigma_j}} e^{-\frac{1}{2(1-r_{ij}^2)} \left( \frac{(z_i-\delta_i)^2}{\sigma_i^2} + \frac{(z_j-\delta_j)^2}{\sigma_j^2} - 2r_{ij} \frac{(z_i-\delta_i)(z_j-\delta_j)}{\sigma_i \sigma_j} \right)}, $$

where $r_{ij}$ is used to quantify the extent to which the two mutation processes are correlated; $\delta_i$ and $\delta_j$ are the mean of expression shifts; $\sigma_i$ and $\sigma_j$ are the corresponding variance of expression shifts. It further assumes that $z_i$’s after different duplication events are independent, e.g. $z_1$ and $z_3$ in figure 3 are independent.
Aside from the mathematical simplicity, the biological significance of this model can be explained as follows. Through gene duplication, the original gene \( k \) becomes two duplicate genes \( i \) and \( j \). On the basis of their ancestral expression \( x_k \), each duplicate gene will alter its expression, summarized as gene shift \( z_i \) (or \( z_j \)), under different selective pressures\(^1\). The Brownian motion \( \sigma^2 t_i \) was used here to account for the neutral differences corresponding to random genetic drift, regardless of its fitness, along branch \( i \). There are two assumptions about independence in the model. One is the independence between the expressions of two duplicate genes \( x \) after a duplication event given their ancestral gene expression; the other is the independence between the expression shifts \( z \) after different duplication events. For the first independence assumption, it is based on the fact that without considering the evolutionary selection force after duplication, the two duplicate gene expressions are involved in two different stochastic processes. For the second independence assumption, the expression shifts refer to all kinds of expression changes except genetic drifts, mostly due to the evolutionary selection pressure; and the expression shifts after different duplication events can be approximately treated as under different selection pressures, therefore, independent.

For the 3-gene phylogeny (figure 3), it can be shown that

\[
(x_1, x_2, x_3)^T \sim \text{Normal} (\mu, V),
\]

where

\[
\mu = \begin{bmatrix}
\mu + \delta_1 + \delta_4 \\
\mu + \delta_2 + \delta_4 \\
\mu + \delta_3
\end{bmatrix},
\]

\[
V = \rho^2 J_3 + \begin{bmatrix}
\sigma_1^2 t_1 + \sigma_4^2 t_4 & \sigma_4^2 t_4 & 0 \\
\ldots & \sigma_2^2 t_2 + \sigma_4^2 t_4 & 0 \\
\ldots & \ldots & \sigma_3^2 t_3 \\
\end{bmatrix}
+ \begin{bmatrix}
\sigma_1^2 s_4 + \sigma_4^2 s_4 + s_4^2 r_{12} & r_{12} s_4 & 0 \\
\ldots & \sigma_2^2 s_4 + s_4^2 & r_{34} s_3 s_4 \\
\ldots & \ldots & s_3^2 \\
\end{bmatrix},
\]

\( J_m \) is a \( m \times m \) matrix of ones.

Note that, no measurement error is included in the model (Equ. 1) because of the lack of replicates in most current microarray data experiments. It can be shown that the

\(^1\) Although we use one variable \( z_i \) to represent the gene shift along branch \( i \), it is indeed the sum of many gene shift events along branch \( i \).
measurement error is independent of the phylogeny of the gene family, and can be estimated based on the replicates and included in the variance-covariance structure, i.e. \( V^* = V + \sigma_e^2 I \), where \( V^* \) is the variance-covariance matrix with the measurement error correction. Here, it is assumed that measurement error is relatively small compared to other variance components in the model, such as \( \rho^2 \) or \( \sigma^2_{\mathbf{J}} \). As the resolution of the microarray technology increases over time, we would expect this assumption to be fair. Later, we also show that, when using the least-squares approach, some of the model parameters are insensitive to the presence of measurement error.

Let \( \mathbf{x} = (x_1, ..., x_n)^T \) be the expression profile of the n-gene family, the above mean and variance-covariance matrices can be generalized to the n-gene case, \( \mathbf{x} \sim N(\mu, V) \), where the elements of \( \mu \) and \( V \) are as follows,

\[
\mu_i = \mu + \sum_{k \in x_i} \delta_k
\]

\[
V_{ij} = \rho^2 + \sum_{m \in x_i, x_j} \sigma^2_{m} t_m + \sum_{h \neq x_i, x_j} s_{h_p} s_{h_q} I_h
\]

(5)

where \( I_h = \begin{cases} 1 & \text{if } i, j \text{ follow the same branch after node } h \\ r_{h_p, h_q} & \text{if } i, j \text{ follow different branches after node } h \end{cases} \)

\( k \in x_i \) refers to all branches along the lineage from root 0 to gene \( x_i \); \( m \in x_i, x_j \) includes all the branches shared by \( x_i \) and \( x_j \) beginning at root 0; \( h \in x_i, x_j \) indexes the set of interior nodes shared by \( x_i \) and \( x_j \), including root 0; \( s_{h_p} \) and \( s_{h_q} \) refer to the branches after node \( h \), which are the same if \( i \) and \( j \) follow the same branch after node \( h \).

In summary, for each element in \( \mu \), there are two components: the mean expression at root 0 (\( \mu \)) and the sum of all the genetic shifts along the evolution lineage of \( x_i (\sum_{k \in x_i} \delta_k) \).

Similarly, each element in \( V \) can be decomposed into the variance at root 0 (\( \rho^2 \)), randomness caused by genetic drifts (\( \sum_{m \in x_i, x_j} \sigma^2_{m} t_m \)) and the variation due to the selection pressure (\( \sum_{h \neq x_i, x_j} s_{h_p} s_{h_q} I_h \)).

---

1 See the "Discussion" sections in chapter II and III.
**Data Structure and Implementation**

Microarray data consist of gene expression measurements under a wide range of experimental conditions. For example, in the "stress" dataset by Gasch (Gasch, A.P. et al. 2000), gene expression levels were measured under 173 different experimental conditions, such as heat shock, hydrogen peroxide treatment, nitrogen depletion, etc. (Table 1 summarizes part of the dataset. Each row refers to the expression profile of one yeast gene, while the columns correspond to different growth conditions, i.e. different microarray chips.) The data measured from each microarray chip is normalized according to the background noise (Shalon et al. 1996; DeRisi et al. 1997), allowing mutual comparison across different experiments/chips.

<table>
<thead>
<tr>
<th>UID</th>
<th>Exp1</th>
<th>Exp2</th>
<th>Exp3</th>
<th>Exp4</th>
<th>Exp5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAL001c</td>
<td>1.53</td>
<td>-0.06</td>
<td>0.58</td>
<td>0.52</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>YAL002w</td>
<td>-0.01</td>
<td>-0.3</td>
<td>0.23</td>
<td>0.01</td>
<td>-0.15</td>
<td></td>
</tr>
<tr>
<td>YAL003w</td>
<td>0.15</td>
<td>-0.07</td>
<td>-0.25</td>
<td>-0.3</td>
<td>-1.12</td>
<td></td>
</tr>
<tr>
<td>YAL004w</td>
<td>0.24</td>
<td>0.76</td>
<td>0.2</td>
<td>0.34</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>YAL005c</td>
<td>2.85</td>
<td>3.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAL006c</td>
<td>-0.22</td>
<td>-0.12</td>
<td>-0.29</td>
<td>-0.51</td>
<td>-0.81</td>
<td></td>
</tr>
<tr>
<td>YAL008w</td>
<td>0.19</td>
<td>0.25</td>
<td>0.69</td>
<td>0.34</td>
<td>0.65</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In equation (5), we showed that under one experimental condition $x \sim N(\mu, V)$. Since we are dealing with several conditions, a subscript is introduced to distinguish different conditions, i.e. $x_g \sim N(\mu, V)$, where $g = 1, ..., G$ for $G$ different conditions, e.g. $g=1$ refers to Exp1 in Table 1. Here, the gene expression profiles of a gene family under different conditions are assumed to follow the same distribution. Assuming independence across different conditions, the joint likelihood can be written as
and the maximum likelihood estimates (MLE's) of parameters can be obtained by applying appropriate numerical algorithms, such as Newton-Raphson or the EM algorithm. Hierarchical likelihood ratio tests (LRT) can be applied for hypothesis testing.

For the general Brownian-based model, there are totally \((7n-5)\) unknown parameters for an \(n\)-gene family, that means even a 3-gene phylogeny will have 16 unknown parameters. To reduce the number of model parameters but still capture some interesting biological information, we proposed a T model, an E model and a C model, all simplified general Brownian-based models. In a derivative model, \(\mu_i\)'s are used without decomposing into \(\mu\) and \(\delta\)'s because the focus of our analysis is the expression variation during evolution.

In the T model, the variation contributed by gene shift is ignored, and so is the expression correlation between progeny genes. Equation (5) is then reduced to

\[
V_y = \rho^2 + \sum_{m \neq s_j, x_j} \sigma^2_m t_m .
\]

When \(t_m\) is known, \(\sigma^2_m\) can be estimated and the hypothesis of constant gene drift rate can be tested using the likelihood method. For the 3-gene family case, the most common null hypothesis is \(H_0: \sigma_1 = \sigma_2\).

In the E model\(^1\), the two gene shifts after gene duplication are assumed to be independent \((r_{ij}=0)\) and equation (6) is re-written as

\[
V_y = \rho^2 + \sum_{m \neq s_j, x_j} (\sigma^2_m t_m + s^2_m) = \rho^2 + \sum_{m \neq s_j, x_j} E_m ,
\]

where \(\sigma^2_m t_m\) and \(s^2_m\) are not identifiable unless information independent of the dataset is provided. The E model provides a general picture of gene expression evolution without knowing \(t_m\), and the whole evolution process can be traced back by looking through those \(E_m\) values, where smaller \(E_m\) values indicate more conservation and larger \(E_m\) values suggest more variation.

Both the T and E models ignore the correlation \((r_{ij})\) between gene shifts \((z_i\) and \(z_j)\) after gene duplication (Equ. 5). To study such correlation, we propose a correlation \((C)\) model where the stochastic variances \((\rho^2, \sigma^2_i\)'s) are assumed to be relatively small compared to the variances \((s^2_i\)'s) associated with dramatic gene shifts after gene duplication. To further

\(^1\) The E model here is the statistical basis for chapters II and III.
simplify the C model, the variances of the two gene shifts after a duplication event are set to be equal. The final formula for the C model can be derived as,

\[ V_{ij} = \sum_{h \cap x_i, x_j} s_h^2 I_h \]  

(7)

where \( I_h = \begin{cases} 1 & \text{if } i, j \text{ follow the same branch after node } h \\ r_h & \text{if } i, j \text{ follow different branches after node } h. \end{cases} \)

Again, \( k \in x_i \) refers to all branches along the lineage from root 0 to gene \( x_i \); \( h \subseteq x_i, x_j \) indexes the set of interior nodes shared by \( x_i \) and \( x_j \), including root 0; \( s_h^2 \) refers to the gene shift variance after node \( h \) and \( r_h \) is the correlation between the two shifts after node \( h \). A major advantage of this C model over previous other models is that the gene expression evolution can be easily characterized by looking at \( r_h \)'s. For example, if \( r_h \) is positive and close to 1, it indicates that no major selection difference between the two duplicates after gene duplication event. On the other hand, if \( r_h \) is close to -1, expressions of both duplicates are under selection but toward different directions. For those \( r_h \)'s close to 0, it is likely that either only one duplicate is under selection pressure or neither duplicate is. The advantages of C model can be summarized as follows. First, since the random stochastic variation is assumed to be negligible, the model no longer needs \( \nu \)'s. Second, unlike in the E model where the comparison of expression variation between different branches depends on two parameters (\( E_i \) and \( E_j \)), one parameter (\( r_h \)) in the C model is enough to represent the expression change after gene duplication.

**Examples**
Figure 3. Phylogenetic inference of the “Glutamyl- and Glutaminyl-tRNA Synthetases” gene family using the Neighbor-joining method.
The gene family "Glutamyl- and Glutaminyl-tRNA Synthetases" (GlnS; figure 3) is used as an example. Proteins of this gene family attach glutamic acid or glutamine to the 3'-terminal ribose residue of its cognate tRNA, forming a glutamyl- or glutaminyl-tRNA. The phylogeny information was obtained from the NCBI website (http://www.ncbi.nlm.nih.gov/cgi-bin/COG/palox?COG0008). There are three baker's yeast (Saccharomyces cerevisiae) genes, YGL245w (gene 1), YOR168w (gene 2) and YOL033w (gene 3) in figure 4, where YOL033w is a mitochondria gene. According to a phylogeny study (Gu 2002), the relative evolutionary time for $t_{ij}$ and $t_{ij}$ is 1.27, $t_3$ is 2.16 and $t_4 = t_3 - t_1 = 0.89$ where the time unit is approximately the divergence time between E. coli and yeast. (In short, the relative evolutionary time is estimated based on the linearized phylogeny of the gene family using the molecular clock approach. See Gu's 2002 for details.) Microarray data is collected from the internet (see reference). Totally, 248 data points ($G=248$) are available for the GlnS gene family.

We first fitted the data using the T model assuming t's are known. The resulting MLE's are

$$\hat{\sigma}_1^2 = 0.206, \hat{\sigma}_2^2 = 0.065, \hat{\sigma}_3^2 = 0.091, \hat{\sigma}_4^2 = 0.644.$$  

Such result shows that, after gene duplication, the evolutionary rates ($\sigma^2$'s) are quite different on the two lineages. Likelihood ratio tests were used to statistically verify the claims. The first hypothesis $H_0$: $\sigma_1^2 = \sigma_2^2$ gave a p-value of 0.0067, suggesting the null hypothesis is not appropriate. Then, the second hypothesis $H_0$: $\sigma_3^2 = \sigma_4^2$ gave a p-value < 0.0001. Both tests
verify our observation that the evolutionary rates are significantly different after gene duplication for the two progeny genes. An interesting fact is that, YOL033w (gene 3), the mitochondria gene, has $\sigma_3^2 = 0.091$ which is much smaller than $\sigma_4^2 = 0.644$. Many references have indicated that the evolution for the mitochondrial genes is more conservative (slow) than genes in the nucleus. Our results coincide with those statements, showing the mitochondrial gene also has smaller evolution variation.

Next, the E model was applied to the same dataset,

$$\hat{E}_1^2 = 0.261, \hat{E}_2^2 = 0.083, \hat{E}_3^2 = 0.196, \hat{E}_4^2 = 0.573,$$

Although no evolutionary time is assumed to be known, which usually is true in practice, we can still have some sense about how much variation is associated with each branch. For example, we observed that $E_4$ has the largest value, which could be explained by either fast evolutionary rate or just long evolutionary time.

Finally, we applied the C model to the data and got

$$r_0 = 0.10, r_4 = 0.72, s_0 = 0.21, s_4 = 0.62.$$

$r_0$ is close to 0, showing that gene expressions diverge after duplication event at root 0. On the other hand, $r_4$ is close to 1, indicating less expression divergence happened between the two lineages after duplication event 4.

**Discussion**

The idea of using a Brownian model for gene expression evolution provides a platform for combining genome sequence information with gene expression information. More complicated models can be achieved by modifying components in the conditional distribution, e.g. Equ. 5 can be modified as $x_k|x_k \sim N(\theta x_k + z_k, \sigma_0^2)$, showing the progeny expression only partially ($\theta$) captures the ancestral gene expression ($x_k$). (However, the biological significance of those complicated models needs to be verified, as does their computational implementation.)
In a typical expression profile data (left), each column \( x_g \) corresponds to the expression measurement of the \( N \) genes at condition \( g \); while the \( n \)-th row \( y_n \) corresponds to the expression of gene \( n \) across all \( G \) conditions. After transformation, \( z_g \)'s are independent across different conditions. \( D \) is the diagonal matrix in spectral decomposition of the variance-covariance matrix of \( y_n \)'s and \( L \) is the corresponding orthologous matrix.

Figure 5. In a typical expression profile data (left), each column \( x_g \) corresponds to the expression measurement of the \( N \) genes at condition \( g \); while the \( n \)-th row \( y_n \) corresponds to the expression of gene \( n \) across all \( G \) conditions. After transformation, \( z_g \)'s are independent across different conditions. \( D \) is the diagonal matrix in spectral decomposition of the variance-covariance matrix of \( y_n \)'s and \( L \) is the corresponding orthologous matrix.

In equation (6), we assume independence across different experimental conditions. For some cases, when the original data are believed to contain several correlated conditions, we provide a strategy which can approximately eliminate the potential correlations among different conditions by data transformation. Suppose the original data is of \( N \) genes under \( G \) experimental conditions (figure 4), where \( x_g \) is the expression profile of \( N \) genes under condition \( g \) and \( y_n \) is the expression profile of gene \( n \) across all conditions. We try to eliminate the dependence among different conditions. Let \( y_n \sim iid N(\mu_n, V) \). (Here, we made a strong assumption that the gene expression profiles of different genes follow the same distribution. To further validate such an assumption, genes that are potentially correlated, such as duplicate genes, can be eliminated from the dataset and the variance estimation is...
computed based on the rest of genes.) Let $\hat{V} = \frac{1}{N-1} \sum_{n=1}^{N} (y_n - \mu_y)(y_n - \mu_y)^T$, where

$\mu_y = \frac{1}{N} \sum_{n=1}^{N} y_n$. Using spectral decomposition, $\hat{V} = LDL^T$, where $D$ is a diagonal matrix of eigenvalues and $L$ is an orthogonal matrix with corresponding eigenvectors. Then, $D^{-1/2}L^{T}y_n \sim N(D^{-1/2}L^{T}H, I)$. The basic strategy here is to estimate $V$ using $y_1, y_2, ..., y_N$; spectral decompose $\hat{V}$ to obtain $L$ and $D$, then transform all $y_n$'s to $D^{-1/2}L^{T}y_n$ (figure 5). In practice, both $N$ and $G$ are usually very large (for yeast, total number of genes $N$ is $\sim 6000$ and the current available microarray conditions $G$ are in hundreds), which makes it possible to result in an approximately singular matrix $\hat{V}$ and the negative eigenvalues because of the potential collinearity between $x_i$'s and the accumulation of the round-up error. One way to deal with negative and very small eigenvalues (since we need to compute $D^{-1/2}$) is to select those eigenvalues which can explain most of the data variation and ignore the rest, then do the transformation.

Throughout this paper, the molecular phylogeny of a gene family is assumed to be given and reliable. Obtaining a reliable phylogeny is not the subject of this paper. However, our model is tree-sensitive because the final multivariate normal distribution depends on the tree topology. Therefore, the most reliable family tree should be used whenever possible. As in all other statistical problems requiring numerical methods to maximize the likelihood, whether the method converges or not is always the key in parameter estimation. We used Newton-Ralphson method to get MLE's, and encountered non-convergence problems from time to time. (The reason might be that our model does not fit the data well.) For the 68 COG gene families with 3 yeast member genes we studied, 7 experienced convergence problems (not shown).

Gene expression is only one measurement for the gene family. There are other aspects interesting to evolutionary biologists. Using the same methodology, proteomic profiles or other continuous measures of genes can be modeled in the same way and the corresponding phylogenetic relationship could be established thereafter. For more discussion about the Brownian-based model, see Gu (2004) for details.
Dissertation Organization

The main focus of my Ph.D. research is to apply statistical modeling to gene expression analysis. By putting gene expression into an evolutionary framework, expression divergence is revealed as expression change after gene duplication, where expression divergence is believed to reflect function divergence between genes. Other issues relevant to the study of gene expression, such as DNA motif structures, are also discussed to demonstrate the limitation of some applications in expression analysis. The core content of this dissertation is composed of three research reports presented as three chapters, each of which is in journal manuscript format. This is followed by a general conclusion chapter which addresses the significance of results we have achieved.

Chapter II is an extension of the method presented in the general introduction chapter by Gu (2004). A conditional distribution approach is used to infer the ancestral expression profiles in the evolution phylogeny. Expression divergence after gene duplication can be directly calculated by comparing expression changes along two lineages after the duplication event. Similar to the sequence divergence analysis, where the divergence can be mapped to certain residues, our expression divergence analysis can map the expression change to some experiment conditions under which expression measurements are taken, revealing possible function roles of duplicate genes as a result of expression divergence.

Chapter III summarizes some techniques in phylogenetic expression analysis. The definition of expression distance is derived from the $E_0$ model of Gu (2004). Since the expression measure is continuous, unlike the sequence data which is discrete, we slightly modify some techniques in molecular phylogeny to fit the expression analysis. Also notice that the handling of the root in the phylogenetic expression analysis is quite different from its counterpart in sequence analysis because of the expression prior associated with the root.

In chapter IV, we explore the relationship between expression profiles and the motif structures in the promoter region. Using the expression data and the motif information of *Saccharomyces cerevisiae*, we demonstrate that such correlation is absent or very weak. Such results coincide with the fact that the expression regulation through the recognition of motif structures is a dynamic process and involves many factors in addition to the one-dimensional motif layout in the upstream region (Wray 2003).
References

Alter, O. et al. (2000) Singular value decomposition for genome-wide expression data processing and modeling. PNAS 97, 10101-10106


Eisen, M.B. et al. (1998) Cluster analysis and display of genome-wide expression patterns. PNAS 95, 14863-14868


Fellenberg, K. et al. (2001) Correspondence analysis applied to microarray data. PNAS 98, 10781-10786
Thomas, J.G. et al. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res. 11, 1227-1236
Thijs, G. et al. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17, 1113-1122
Wen, X. et al. (1998) Large-scale temporal gene expression mapping of central nervous system development. PNAS 95, 334-339

On-line microarray datasets:
http://www.genome.stanford.edu/yeast_stress
http://cellcycle-www.stanford.edu
http://genome-www.stanford.edu/zinc
CHAPTER II: A FAST METHOD FOR INFERRING ANCESTRAL EXPRESSION STATES AFTER GENE DUPLICATIONS

A paper to be submitted to Genetics

Zhongqi Zhang and Xun Gu

Abstract

Function divergence after gene duplication has been considered to be an important mechanism for the evolution of new functions. Although gene expression profiles have been treated as an important indicator of gene function, large scale gene expression analysis has mostly focused on current relationships among genes, instead of their evolutionary relationships. By putting expression analysis into the framework of evolution, we make inferences about expression divergence after gene duplication. Based on the Brownian-based model (Gu 2004), the posterior distribution of the ancestral expression profiles are shown to follow a multivariate-normal distribution. This approach provides not only the estimates of the ancestral expression profiles, but also provides a measure of the precision of the estimation/prediction, thereby, filtering significant information from the background noise of the data.

Introduction

Ancestral state reconstruction within an evolutionary tree is at the center of comparative studies in evolutionary biology. At the morphological level, comparative analysis looking for evidence of correlated change in two characters may need the help of inferred ancestral states (Harvey and Pagel, 1991). At the molecular level, the properties of ancient molecules (Malcolm et al., 1990; Stackhouse et al., 1990; Adey et al., 1994; Jermann et al., 1995) can be examined through the inferred ancestral amino acid sequences, which can be further tested in vivo or in vitro. Massive microarray expression profiles make it possible to reconstruct the ancestral expression pattern. In the course of such endeavor, appropriate methodology for ancestral state inference is essential.
Among several methods\textsuperscript{1} proposed to reconstruct the ancestral states on a
phylogenetic tree, the parsimony method (Eck and Dayhoff 1966; Fitch 1971; Hartigan 1973)
is probably the oldest and the most frequently used in comparative studies. Recently,
statistically sound approaches have been developed to reconstruct DNA, amino acid, and
discrete morphological data (e.g., Schluter, 1995; Yang et al., 1995; Schluter et al., 1997;
Pagel, 1999). The maximum likelihood method finds the character states at the interior nodes
of the tree that maximize the probability of observing the data (e.g., Pagel 1999). The goal of
ancestral inference, on the other hand, is to calculate the posterior probability that an
ancestral node on a tree has a particular state, given the observations at the tips (exterior
nodes) of the tree. The posterior analysis depends on several factors: (1) the topology of the
phylogenetic tree which shows the relationship among genes, both current genes and
ancestral genes; (2) branch lengths which measure to what extent genes are related; (3)
parameters of the evolutionary model, which are estimated using the maximum likelihood
approach in the case of the likelihood model.

Many comparative methods are designed specifically for categorical or discrete
variables. We have recognized that ancestral estimation has also been important in the study
of evolutionary genomics, where the characters could be continuous, e.g., the microarray
expression profile. The most common way of estimating the ancestral states of continuous
characters was to use the principle of parsimony. For instance, the sum of squared changes
parsimony algorithm (Huey and Bennett, 1987; Maddison, 1991; McArdle and Rodrigo,
1994) estimates the phenotype of each ancestor as a weighted average of all the phenotypes
measured for exterior species. The weights used in calculating averages correspond to the
phylogenetic distances between extant species and the ancestor being estimated. Schluter et
al. (1997) use a maximum likelihood approach to estimate ancestral states, which is
essentially to minimize the sum of squared changes under a Brownian model similar to
Felsenstein’s model (1985). In this article we propose a simple approach based on the
conditional distributions of ancestral states given current states to infer ancestral states of
continuous molecular characters. Potential applications for comparative genomics are
illustrated by examples.

\textsuperscript{1} For a detail description of many methods for reconstructing phylogenetic trees, see the introduction section in chapter III.
Ancestral Expression Inference

For microarray data, the expression level \( x \) of a gene is measured by the log-transformed signal intensity after normalization and bias-correction (Quackenbush 2001). For an \( n \)-member gene family with a given (rooted) phylogenetic tree, let \( x = (x_1, ..., x_n)^T \) be the observed expression pattern, and \( y = (y_1, ..., y_m)^T \) be the ancestral expressions at ancestral (interior or root) nodes\(^1\). (For a bifurcating phylogeny, \( n \) exterior nodes correspond to \((n-1)\) interior nodes.) According to the Bayes rule, the posterior density \( P(y_1, ..., y_m | x_1, ..., x_n) \), concisely written as \( P(y | x) \), can be computed as follows

\[
P(y | x) = \frac{P(x, y)}{P(x)},
\]

where \( P(x, y) \) is the joint density of \( x \) and \( y \) and \( P(x) \) is the “marginal” density of current expression. The posterior mean of \( y \) is \( \mu_{yx} = E[y|x] = \int y P(y|x) \, dy \) and the posterior variance of \( y \) is \( V_{yx} = E[yy^T|x] = E[y|x]E[(y|x)]^T \), which measures the accuracy of estimation/prediction.

---

\(^1\) \( x \) and \( y \) refer to the expression profiles under one experimental condition.
Under the Eo model (Gu 2004), it has been shown that, for an n-member gene family with a given (rooted) phylogenetic tree, the joint density of gene expressions \( P(x) = P(x_1, ..., x_n) \) follows a multivariate normal distribution \( N(x; \mu, V) \), where \( \mu_i = E(x_i) \) and

\[
V_{ij} = \rho^2 + \sum_{k \in x_i, x_j} E_k; \quad k \in x_i, x_j \text{ includes all branches shared by } x_i \text{ and } x_j \text{ since root 0 and } E_k \text{ is the sum of variances along branch } k.
\]

Let \( M = m + n \), where \( m \) is the number of interior (ancient) nodes and \( n \) is the number of exterior (current) nodes. Following the derivation formulated by Gu (2004), it has been shown that \( P(y, x) \) is an M-variate normal density, with \( M \times M \) variance-covariance matrix \( V_M \). For notational convenience, the M-variate normal density is written in the matrix format as,

\[
\begin{bmatrix} y \\ x \end{bmatrix} \sim N \left( \begin{bmatrix} \mu_y \\ \mu \end{bmatrix}, V_M = \begin{bmatrix} A & H' \\ H & V \end{bmatrix} \right)
\]

(2)

where each element of \( \mu_y \) is the prior mean at root 0; the \( ij \)-th element of the \( m \times n \) matrix \( H \), \( H_{ij} = \rho^2 + \sum_{k \in y_i, y_j} E_k \), is the expression covariance between the ancestral node \( y_i \) and the exterior (current) node \( x_j \), and the \( ij \)-th element of the \( m \times m \) matrix \( A \), \( A_{ij} = \rho^2 + \sum_{k \in y_i, y_j} E_k \), is the expression covariance between the \( i \)-th and \( j \)-th ancestral nodes. For example, figure 1 is a typical rooted three-gene family, where \( x = (x_1, x_2, x_3) \), \( y = (y_4, y_0) \) and \( M = m + n = 2 + 3 = 5 \).

Since \( P(x, y) \) has multi-normal density, the posterior density of ancestral nodes \( y \),

\[
P(y | x) = \phi((x', y''); \mu_M, V_M) / \phi(x; \mu, V) \]

also has an m-variate normal density, written as \( \phi(y | x; \mu_{y|x}, V_{y|x}) \), where \( \mu_{y|x} = (\mu_{y1|x}, ..., \mu_{ym|x})' \) is the posterior mean vector of the ancestral nodes, and \( V_{y|x} \) is the \( m \times m \) posterior variance-covariance matrix of \( y_1, ..., y_m \) given \( x \). The analytical formulas are

\[
\begin{align*}
\mu_{y|x} &= \beta_0 + HV^{-1}x \\
V_{y|x} &= A - HV^{-1}H'
\end{align*}
\]

(3)

where \( \beta_0 = \mu_y - HV^{-1} \mu \), which does not depend on \( x \).

---

1 Also see chapter 1 for a brief review.
2 \( k \in x_i, y_j \) refers to all branches shared by \( x_i \) and \( y_j \) since root 0.
3 \( k \in y_i, y_j \) refers to all branches shared by \( y_i \) and \( y_j \) since root 0.
Using the maximum likelihood (ML) approach\(^1\) (Gu 2004), model parameters are estimated, \(\hat{\rho}^2, \hat{E}_1, \ldots\); and used to construct \(\hat{\mu}_M, \hat{V}_M\). For the three-gene family case, we have

\[
\hat{\mu}_M = \begin{bmatrix}
\hat{\mu}_4 \\
\hat{\mu}_1 \\
\hat{\mu}_2 \\
\hat{\mu}_3
\end{bmatrix}, \quad \hat{V}_M = \begin{bmatrix}
\hat{\rho}^2 + \hat{E}_4 & \hat{\rho}^2 & \hat{\rho}^2 + \hat{E}_4 & \hat{\rho}^2 + \hat{E}_4 & \hat{\rho}^2 \\
\vdots & \hat{\rho}^2 & \vdots & \hat{\rho}^2 & \vdots & \hat{\rho}^2 \\
\hat{\rho}^2 + \hat{E}_4 + \hat{E}_4 & \hat{\rho}^2 + \hat{E}_4 & \hat{\rho}^2 & \hat{\rho}^2 & \hat{\rho}^2 \\
\vdots & \vdots & \hat{\rho}^2 + \hat{E}_2 + \hat{E}_4 & \hat{\rho}^2 & \hat{\rho}^2 & \hat{\rho}^2 \\
\vdots & \vdots & \vdots & \hat{\rho}^2 + \hat{E}_3 & \hat{\rho}^2 & \hat{\rho}^2
\end{bmatrix}.
\]

(4)

Then \(\mu_{y|x}\) and \(V_{y|x}\) can be computed accordingly using Equ. 3.

**Fast Algorithm for Ancestral Expression Inference**

To infer the ancestral expressions of a gene family by the posterior distribution, one must know the structure of the phylogenetic tree, the three matrices, \(V, H, A\), and two vectors, \(\mu, \mu\). Then, the posterior inference can be made through equation (3). The likelihood approach is sensitive to model assumption, i.e. convergence problems may arise during numerical estimation when the data do not satisfy the model assumptions. For example, in Equ. 4, \(\hat{\rho}^2\) is the variance of \(y_4\) which has to be positive, but using the ML approach \(\hat{\rho}^2\) can end up with a negative value. Constraint or transformation can be applied to \(\rho^2\) during numerical estimation to ensure its positive value, but one usually encounters convergence problem thereafter. To overcome such difficulty and provide a convenient computational approach, we propose a fast algorithm which can be easily implemented without using the likelihood to estimate model parameters. This algorithm is equivalent to solving a set of linear equations using the least-squares (LS) approach (see Appendix). Although LS estimators will lose efficiency compared to ML estimators, this fast algorithm is robust against violations of model assumptions. Later, simulation analysis shows that the efficiency loss seems worth the gain in robustness. The detailed steps of this algorithm are summarized as follows:

\(^1\) Also see chapter I for a brief review.
(1) Phylogenetic inference and data preparation:

The first step is to infer the phylogeny of the gene family from amino acid or DNA sequences, which can be performed by any conventional approach, such as neighbor-joining (NJ), parsimony (MP), or likelihood (ML)\(^1\). An estimate of the expression variance-covariance matrix \( \hat{V} \) is calculated from an array \((G)\) of microarray expression data. Let \( \mathbf{x}_i = (x_{i1}, ..., x_{ig})' \) be the expression profiles for gene \( i \) under \( G \) (independent) experimental conditions, the (sample) mean of gene \( i \) is

\[
\bar{x}_i = \frac{1}{G} \sum_{g=1}^{G} x_{ig}, \quad i = 1, ..., n. \tag{5}
\]

and the estimate of the \( ij \)-th element of matrix \( (V) \) can be computed as

\[
\hat{V}_{ij} = \frac{\sum_{g=1}^{G} (x_{ig} - \bar{x}_i)(x_{ig} - \bar{x}_j)}{(G-1)} \tag{6}
\]

Here, we assume \( x_g = (x_{1g}, x_{2g}, ..., x_{ng})' \), the expression profiles of the \( n \)-gene family under condition \( g \) are i.i.d. with \( N(\mu, \Sigma) \).

(2) Matrices \( \mathbf{H} \) and \( \mathbf{A} \):

Though the matrices \( \mathbf{H} \) and \( \mathbf{A} \) cannot be estimated directly from the microarray data, we found a simple solution under the \( E_0 \) model (Gu 2004) using the concept of expression distance. As shown in the appendix, the following estimates are indeed equivalent to the least-squares (LS) estimates when solving a set of linear equations based on the \( E_0 \) model.

We start with matrix \( \mathbf{H} \). For any ancestral node \( i \) and the current node \( j \), let \( a_{ij} \) be the most recent common ancestor (MRCA) of \( i \) and \( j \). By definition, \( a_{ij} = i \) if node \( i \) is the ancestor of the current node \( j \), e.g. in figure 1, node 4 is the MRCA of \( i=4 \) and \( j=1 \); node 0 is the MRCA of \( i=4 \) and \( j=3 \). Let \( C_1 \) and \( C_j \) be the two clusters of exterior taxa after the bifurcate duplication event after node \( a_{ij} \), where \( C_j \) is the cluster containing gene \( j \), e.g. in figure 1, for \( a_{41} \), \( C_4 = \{2\} \) and \( C_1 = \{1\} \); for \( a_{03} \), \( C_0 = \{1,2\} \) and \( C_3 = \{3\} \); for \( a_{43} \), \( C_4 = \{1,2\} \) and \( C_3 = \{3\} \). Then, it can be shown that \( H_{ij} \) is the average covariance between gene \( j \) and genes in cluster \( C_1 \) (see Appendix).

\(^1\) For detail information about tree making methods, see “Molecular Evolution and Phylogenetics” by Nei and Kumar, 2002; or see the introduction section of chapter III for a brief review.
\[
H_{ij} = \frac{\sum_{k \in C_i} \sum_{l \in C_j} V_{kj}}{c_i}
\]

where \( c_i \) is the number of genes in cluster \( C_i \). For example, \( H_{41} = V_{12} \); \( H_{03} = (V_{13} + V_{23})/2 \).

Next we consider matrix \( A \). Similar to the above, for any two ancestral nodes \( i \) and \( j \), let \( a_{ij} \) be the most recent common ancestor (MRCA) of \( i \) and \( j \). Again, let \( C_1 \) and \( C_j \) be the two exterior gene clusters after node \( a_{ij} \), the \( ij \)-th element of \( A \) is given by

\[
A_{ij} = \frac{\sum_{k \in C_i} \sum_{l \in C_j} V_{kl}}{c_i c_j}
\]

where \( c_i \) and \( c_j \) are the number of genes in \( C_1 \) and \( C_j \). For example, \( A_{40} = (V_{13} + V_{23})/2 \); \( A_{44} = V_{12} \).

(3) Vectors \( \mu \) and \( \mu_y \):

For \( \mu \), the mean expression for the exterior genes \( x \), the sample mean (Equ. 5) is used as an estimate\(^1\). The estimate of \( \mu_y \), the mean expression for the interior genes \( y \), is based on the Brownian-based model (Gu 2004), where the expression difference for different genes across different conditions is modeled solely as a result of stochastic error – Brownian motion. Under this assumption, the elements of \( \mu_y \) are estimated as the grand mean expression\(^2\) of all exterior genes across all experimental conditions in the dataset, i.e.

\[
\hat{\mu} = \frac{1}{n \times G} \sum_{i=1}^{n} \sum_{g=1}^{G} x_{ig}
\]

Then, the estimates of the model parameters can be written analytically using equation (5) – (9). For the three-gene family in figure 1, Equ. 4 can be written analytically as

\(^1\) In both cases, MLE can give more efficient estimates. However, unlike MLE which is sensitive to the model assumption, LS estimation we used here is robust against the violation of model assumptions.
\[
\hat{\mu}_M = \begin{bmatrix}
\hat{\mu} \\
\hat{\mu} \\
\bar{x}_1 \\
\bar{x}_2 \\
\bar{x}_3 \\
\end{bmatrix}, 
\hat{V}_M = \begin{bmatrix}
\hat{V}_{12} & \frac{\hat{V}_{13} + \hat{V}_{23}}{2} & \hat{V}_{12} & \frac{\hat{V}_{13} + \hat{V}_{23}}{2} \\
\frac{\hat{V}_{13} + \hat{V}_{23}}{2} & \hat{V}_{12} & \frac{\hat{V}_{13} + \hat{V}_{23}}{2} & \hat{V}_{13} \\
\frac{\hat{V}_{13} + \hat{V}_{23}}{2} & \hat{V}_{12} & \hat{V}_{23} & \hat{V}_{13} \\
\frac{\hat{V}_{13} + \hat{V}_{23}}{2} & \hat{V}_{13} & \hat{V}_{23} & \hat{V}_{33} \\
\end{bmatrix}
\tag{10}
\]

which is equivalent to the Equ. 4 but using the least-squares estimates (see Appendix for proof).

**Implementation**

During implementation, $\hat{V}_M$ requires special attention (Equ. 10) because it may not be non-negative definite under a few cases. (Equ. 3 holds when $V_M$ is positive definite, for example, but $V_{y|x}$ may not be non-negative definite if $V_m$ is not non-negative definite.) To ensure the resulting posterior variance $V_{y|x}$ is nonnegative (or even positive) definite, some manipulation needs to be done with $V_M$ when it is not non-negative definite and one of the common strategies is to decompose $\hat{V}_m$ by spectral decomposition,

\[
\hat{V}_M = P \hat{D} P' = \begin{bmatrix}
p_1 \\
p_2 \\
... \\
p_M \\
\end{bmatrix} \begin{bmatrix}
\hat{\lambda}_1 \\
\hat{\lambda}_2 \\
... \\
\hat{\lambda}_M \\
\end{bmatrix} \begin{bmatrix}
p_1' \\
p_2' \\
... \\
p_M' \\
\end{bmatrix}
\tag{11}
\]

where $\hat{\lambda}_k$ is the eigenvalue and $p_k$ is the eigenvector associated with $\hat{\lambda}_k$. Negative eigenvalues are set to zero (or some very small number if desire positive definite), resulting in $D^*$, and the nonnegative definite matrix $\hat{V}_M^*$ is restored by $P \times D^* \times P'$, which usually only slightly deviates from $\hat{V}_M$.

We have developed a C++ computer program to infer the ancestral states of gene expressions based on the fast algorithm. As schematically described in Fig. 2, the input of amino acid or DNA sequence alignment is for phylogenetic inference. (In the current version, the neighbor-joining method is used.) After inputting expression profiles of the gene family,
one can infer ancestral expression profiles for each internal node by the posterior means, and the precision of ancestral prediction is measured by the posterior variance (Equ. 3).

**Input:**
- dataset of sequences
- dataset of gene expression profiles

**Phylogenetic inference:**
- NJ phylogeny inference
- Root identification

**Estimation:**
- compute sample means $x_i$ across G conditions to estimate $\mu$
- compute grand mean of $x_i$’s to estimate $\mu_y$
- compute sample variance/covariance $\hat{V}_{ij}$
- estimate $H$ and $A$, then construct $\hat{V}_M$
- if $\hat{V}_M$ not non-negative definite, compute the spectral decomposition of $\hat{V}_M$ and set negative eigenvalues to 0

**Posterior inference:**
- posterior variance: $\hat{V}_{y|x}$
- for expression profile $x$ of the exterior genes at each condition $g$ {
  - compute posterior mean of interior genes at condition $g$: $\hat{\mu}_{y|x}$
}

Figure 2. The ancestral expression inference.

**An Example: Ancestral Expression Inference for Yeast Enolase gene family**

The enolase gene family (Eno) of baker’s yeast, Saccharomyces cerevisiae, has five member genes, YGR254w, YHR174w, YMR323w, YOR393w and YPL281c. (Enolase is
one of the key enzymes in both Glycolysis and Gluconeogenesis, catalyzing the reversible reaction between 2-phosphoglycerate (2PG) and phosphoenolpyruvate (PEP). YGR254w is also known as enolase I (ENO1) and YHR174w as enolase II (ENO2). The other three genes are hypothetical ORFs with sequences similar to Eno, but with unknown functions. The protein sequence alignment file was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/COG/aln/COG0148.aln) and the phylogenetic tree was made using the neighbor-joining method (NJ) in MEGA2 (http://www.megasoftware.net/; figure 3a). The rooted phylogeny of the five yeast enolases is diagrammed in figure 3b.
Figure 3. (a) Neighbor-joining (NJ) tree of enolase gene family (COG0148). The highlighted branches are the five-gene tree of *Saccharomyces cerevisiae*. (b) The schematic diagram (topology only) of 5-yeast-gene family. The root is referred as 0.
Microarray data were collected from on-line resources (see reference 27). After ancestral expression analysis, gene expression of gene 6, 7, 8 and 0 can be written as,

\[
\mathbf{y}_g | \mathbf{x}_g = \begin{bmatrix} y_{6g} \\ y_{7g} \\ y_{8g} \\ y_{0g} \end{bmatrix} \sim N(\mu_{y|x}, \mathbf{V}_{y|x}),
\]

where \( \mu_{y|x} = \begin{bmatrix} 0.076 \\ 0.081 \\ -0.056 \\ 0.008 \end{bmatrix} + \begin{bmatrix} 0.168 \\ 0.156 \\ -0.048 \\ -0.233 \end{bmatrix} \mathbf{x}_g \)

\( \mathbf{V}_{y|x} = \begin{bmatrix} 0.053 & 0.056 & -0.002 & -0.014 \\ 0.056 & 0.059 & -0.002 & -0.015 \\ -0.002 & -0.002 & 0.044 & -0.004 \\ -0.014 & -0.015 & -0.004 & 0.004 \end{bmatrix} \)

\( g \) stands for the experiment condition, \( g = 1, \ldots, G \) and for the Eno family data, we have \( G=230 \). Note that, the mean ancestral expression depends on the current expression profile at condition \( g \), while the ancestral variance does not.

Based on the posterior distribution, we can further explore the expression changes along the gene family evolution. For example, we are interested in how expression profiles diverged after the duplication event at node 8 (figure 3(b)). The expression difference between gene4 (ENO1) and its ancestral gene 8 has the form,

\[
x_{4g} - y_{8g} | x = x_1 g + x_2 g - x_3 g - x_5 g
\]

where \( \hat{x}_{4g} - \hat{x}_{8g} | x = 0.056 + 0.048 \cdot x_{1g} - 0.016 \cdot x_{2g} - 0.051 \cdot x_{3g} + 0.195 \cdot x_{4g} - 0.113 \cdot x_{5g} \)

and \( \hat{V}_{y|x} = 0.044 \).

Then, the expression change is statistically significant if \( \frac{|x_{4g} - y_{8g}|}{0.044} \geq \phi_{\alpha/2} \), where \( \alpha \) is the significant level and \( \phi_{\alpha/2} \) is the quantile of standard normal for percentile \( \alpha \). Usually, the expression changes are calculated for all \( G \) conditions and the issue of multiple comparison

\(^1\) 0.044 is used to approximate the prediction variance.
should be taken into account. Here, the Bonferroni method was used, resulting in

\[ \frac{|x_{4g} - y_{4g}|}{\sqrt{0.044}} \geq \frac{1}{\alpha} \cdot 1 - \frac{1}{2G} \].

Figure 4(a) shows the expression change of YGR254w (ENO1) after gene duplication at node 8. None of those differences is significant at the 0.05 confidence level, suggesting no expression change for ENO1 after its duplication. Figure 4(b) is the bar-plot of expression change for YHR174w (ENO2) and 13 changes are significant at the 0.05 level. Among those 13 significant expression changes, 12 show a decrease after duplication and 9 of them are in the sporulation experiment (Chu et al. 1998). Those results suggest that, the function divergence after gene duplication between ENO1 and ENO2 probably occurs in the ENO2 lineage, especially under the condition of sporulation where the expression of ENO2 is repressed significantly. In another word, after the gene duplication between ENO1 and ENO2, the evolution constraint along the ENO2 lineage is relaxed due to the function redundancy from ENO1 which retains most of the ancestral function; while ENO2 diverges from its ancestral function.

![Figure 4](image-url)

Figure 4. Bar-plots of gene expression change after gene duplication. (a) Expression change of YGR254w (ENO1) after its most recent duplication (Figure 3). (b) Expression change of YHR174w (ENO2). The dash lines indicate simultaneous 95% prediction intervals.
Discussion

Using the ancestral expression inference approach, the expression divergence after each duplication event across the whole phylogeny can be estimated. Since the expression profile can be used as an indicator of the function activity profile, the analysis of expression divergence provides insight about the function divergence between duplicate genes after duplication. For some cases, when the relative evolutionary time for each evolution path is available (usually from the molecular phylogenetic analysis), the expression evolution rates can also be calculated along all evolution paths, showing a general picture of expression divergence rate for the whole family.

Effect of measurement error

In the Brownian-based model $E_0$ model (Gu 2004), the measurement error is assumed to be negligible; its variance is assumed to be small compared to the variance components in the model. We show that, if the measurement error is not negligible, the least-squares approach still gives unbiased estimation for matrices $A$, $H$ and the off-diagonal elements in matrix $V$; while the estimates of diagonal elements of matrix $V$ are biased.

Suppose $x_{ig}(obs) = x_{ig} + e_{ig}$ and $e_{ig}$ i.i.d. with variance $\sigma_e^2$. It can be shown that

$$E(V_{\hat{}}_i) = \begin{cases} V_{ij} + \sigma_e^2 & \text{if } i = j \\ V_{ij} & \text{if } i \neq j \end{cases}$$

Since the elements of $H$ (Equ. 7) and $A$ (Equ. 8) only depend on the covariance between expressions of taxa, they are unbiased under the least-squares estimation procedure.

Least-squares (LS) vs. maximum likelihood (ML)

The fast algorithm of constructing matrix $V_M$ (Equ. 2) is based on the least-squares (LS) approach to solve a set of linear equations (Equ. a6 in Appendix). This approach is robust against violations of model assumptions in estimating matrices $H$ and $A$ when the measurement error is non-negligible. There are other approaches which are more efficient in estimating model parameters, such as the maximum likelihood method (ML); however, most

---

1 $x_{ig}$: The gene expression measurement of gene i under condition g.
of these approaches require numerical iterations to find the estimators, suffering severely for the violation of model assumptions. For example, when using ML to estimate model parameters in E0 model (Gu 2004), the estimate of the variance-covariance matrix has to be positive-definite, otherwise, the numerical algorithm will fail to converge. To evaluate the performance of LS, we did some simulation work to compare it with the ML approach.

First, we study the efficiency of the LS approach (assuming there is no measurement error or measurement error is negligible) compared with ML approach. Based on a three-gene family E0 model, our simulation results (Table 1) show that the relative efficiency\(^1\) between LS and ML is between 0.85-0.95, about 10% efficiency is lost by using the LS approach. However, its robustness against deviation from model assumptions (see next paragraph) and easy computation make it worthy consideration, in spite of the loss in efficiency.

Table 1(a). Relative efficiency\(^2\) between LS and ML approaches.

<table>
<thead>
<tr>
<th></th>
<th>(\rho^2) (0.03)*</th>
<th>E1 (0.26)</th>
<th>E2 (0.08)</th>
<th>E3 (0.2)</th>
<th>E4 (0.57)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No (e_{ig})</td>
<td>0.92</td>
<td>0.93</td>
<td>0.96</td>
<td>0.93</td>
<td>0.86</td>
</tr>
<tr>
<td>(e_{ig} \sim N(0, 0.001))</td>
<td>0.93</td>
<td>0.89</td>
<td>0.90</td>
<td>0.92</td>
<td>0.90</td>
</tr>
</tbody>
</table>

*: the values in parentheses refer to the true parameter values, which are from the analysis with COG0008\(^3\).

Table 1(b). Simulation results when there is no measurement error.

<table>
<thead>
<tr>
<th></th>
<th>(\rho^2) (0.03)*</th>
<th>E1 (0.26)</th>
<th>E2 (0.08)</th>
<th>E3 (0.2)</th>
<th>E4 (0.57)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML Mean*</td>
<td>0.0294</td>
<td>0.230</td>
<td>0.0783</td>
<td>0.199</td>
<td>0.569</td>
</tr>
<tr>
<td>Variance</td>
<td>0.000767</td>
<td>0.00149</td>
<td>0.00106</td>
<td>0.00121</td>
<td>0.00417</td>
</tr>
<tr>
<td>LS Mean</td>
<td>0.0296</td>
<td>0.0259</td>
<td>0.0784</td>
<td>0.200</td>
<td>0.569</td>
</tr>
<tr>
<td>Variance</td>
<td>0.000833</td>
<td>0.00160</td>
<td>0.00110</td>
<td>0.00130</td>
<td>0.00485</td>
</tr>
</tbody>
</table>

*: mean and variance is calculated based on simulation results.

\(^1\) Relative efficiency: \(r.e. = \frac{\text{Var}_{ML}(\theta)}{\text{Var}_{LS}(\theta)}\).
\(^2\) The result is based on 2000 simulated datasets, with sample size 200.
\(^3\) See chapter I "Examples" for details.
Next, we study the effect of measurement error in the LS and ML approaches. When measurement error variance is relatively small, e.g. 10% of the smallest variance component in the model parameters, the performances of both LS and ML are almost the same as if there is no measurement error and there is no convergence problem. (The performance, here, refers to the relative efficiency between the two approaches.) When the measurement error is on the same scale as the small variance components in the model, in about 1 out of 5 cases, the ML approach does not converge due to either negative determinant or singularity of the estimated variance-covariance matrix. If the measurement error variance is further increased to the same scale as the larger variance components in the model, the chance of encountering a convergence problem in the ML approach increases to as high as 50%. These results reveal that, although the ML approach tends to provide more efficient estimators than the LS approach, its performance can be highly deteriorated by the presence of measurement error. We, therefore, suggest using the LS approach in the current expression phylogenetic study instead of the ML approach.

Another advantage of the LS approach over the ML approach is its efficiency in computational time. In the ML approach, the computation of the likelihood requires the calculation of matrix determinant and matrix inverse which will take much of the computational time, especially when the size of the gene family is large and the numerical algorithm undergoes lots of iterations before convergence.

References


27. On-line microarray datasets:
http://www.genome.stanford.edu/yeast_stress
http://cellcycle-www.stanford.edu

http://genome-www.stanford.edu/zinc


Appendix

$P(y, x)$ is a multivariate normal density under the $E_0$ model

Let $M = n + m$. According to the $E_0$-model we developed (Gu 2004), given an ancestral expression $a$, the gene expression after $t$ time units follows a normal distribution $B(x|a) = N(x; a, E)$, where $E$ is the sum of expression variation along the branch. When the phylogenetic tree is known, under the assumption of independent evolution among branches, the joint density of $z = (y', x')' = (y_1, ..., y_{n+m}, x_1, ..., x_n)'$, conditional on root expression value $y_0$, is simply given by

$$P(z|y_0) = \prod_{a \in \{y, O\}} B(z_a|a)$$

where $\{a \in y, O\}$ runs over all ancestral expressions including root $O$; $z_a \in x$ or $y$ is the end-node corresponding to the ancestral node $a$. In other words, $P(z|y_0)$ is a product of normal densities along all branches. For instance, the joint density for 4-member genes (figure 1a) is given by

$$P(z|y_0) = B(y_5|y_0)B(y_6|y_0)B(x_1|y_5)B(x_2|y_5)B(x_3|y_6)B(x_4|y_6)$$

Consequently, $P(z|y_0)$ is a multi-normal density. Moreover, if the prior of the root, $\pi(y_0)$, is normal, one can show that $P(z) = \int P(z|y_0) \pi(y_0) dy_0$ is a $M \times M$ multivariate normal
density. Following the derivation formulated by Gu (2004), the $M \times M$ variance-covariance matrix $V_M$ for $z$ can be written as follows

$$V_M = \begin{bmatrix} A & H \\ H' & V \end{bmatrix}$$  \hspace{1cm} (a3)$$

where $H (n \times m)$ is the covariance matrix between ancestral genes ($y$) and current genes ($x$); $A (m \times m)$ is the variance-covariance matrix for the ancestral genes and $V$ is the variance-covariance matrix of current genes.

Under the $E_0$ model, the $ij$-th element of $H$ is given by

$$H_{ij} = \rho^2 + \sum_{k \in x_i, y_j} E_k, \quad i = 1, ..., n; \quad j = n+1, ..., n+m$$  \hspace{1cm} (a4)$$

where $k \in x_i, y_j$ runs over all branches shared by the current expression $x_i$ and the ancestral expression $y_j$ since root $O$. The matrix $A$ is the variance-covariance matrix among ancestral nodes and its $ij$-th element is given by

$$A_{ij} = \begin{cases} \rho^2 + \sum_{k \in y_i, y_j} E_k & \text{if } i = j \\ \rho^2 + \sum_{k \in y_i, y_j} E_k & \text{if } i \neq j \end{cases}, \quad i, j = n+1, ..., n+m$$  \hspace{1cm} (a5)$$

where $k \in y_i$ runs all branches from root $O$ to $y_i$ and $k \in y_i, y_j$ refers to all branches shared by $y_i$ and $y_j$ since root $O$.

**Computation of $H$ and $A$ by expression distances**

We have proposed an evolutionary model ($E_0$) for gene expression divergence of a gene family (Gu 2004). Under the $E_0$ model, Gu (2004) has shown that, for an $n$-member gene family with a given (rooted) phylogenetic tree, the joint density of gene expressions $P(x) = P(x_1, ..., x_n)$ follows a multi-variate normal distribution $N(x_1, ..., x_n; \mu, V)$. The mean vector is $\mu = (\mu, ..., \mu)'$ and the $ij$-th element of the variance-covariance matrix $V$ is given by

$$V_{ij} = \begin{cases} \rho^2 + \sum_{k \in x_i, x_j} E_k & \text{if } i = j \\ \rho^2 + \sum_{k \in x_i, x_j} E_k & \text{if } i \neq j \end{cases}$$  \hspace{1cm} (a6)$$

where $E_k$ is called the expression length of branch $k$; $k \in x_i$ 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 $O$. 
For an n-gene family, there are \((n^2+n)/2\) different \(V_{ij}\)'s \(V_{ij} = V_{ji}\) and \((2n-1)\) model parameters including \(\rho^2\). For \(n \geq 2\), there are more equations than the number of parameters and parameters can be estimated using a least-squares approach. For the 3-gene family in figure 1, we have

\[
\begin{align*}
V_{11} &= \rho^2 + E_1 + E_4 \\
V_{12} &= \rho^2 + E_4 \\
V_{13} &= \rho^2 \\
V_{22} &= \rho^2 + E_2 + E_4 \\
V_{23} &= \rho^2 \\
V_{33} &= \rho^2 + E_3 \\
\end{align*}
\]

(a7)

The least-squares estimators\(^1\) are

\[
\begin{align*}
\hat{\rho}^2 &= \frac{\hat{\rho}_{13} + \hat{\rho}_{23}}{2} \\
\hat{E}_1 &= \hat{V}_{11} - \hat{V}_{12} \\
\hat{E}_2 &= \hat{V}_{22} - \hat{V}_{12} \\
\hat{E}_3 &= \hat{V}_{33} - \frac{\hat{V}_{13} + \hat{V}_{23}}{2} \\
\hat{E}_4 &= \hat{V}_{12} - \frac{\hat{V}_{13} + \hat{V}_{23}}{2} \\
\end{align*}
\]

(a8)

where \(\hat{V}_{ij}\) is as in Equ. 6. After estimating \(\rho^2\) and the \(E_i\)'s explicitly, the matrices \(A\) and \(H\) are estimated by substituting those estimates into Equ. a4 and Equ. a5.

\(^1\) Here, we used a fast algorithm to analytically solve the linear equations using least-squares without using the normal equations. (See Appendix of chapter III for details.)
CHAPTER III: RECONSTRUCTING THE PHYLOGENETIC TREES FROM GENE EXPRESSION DATA USING THE DISTANCE METHODS

A paper to be submitted to *Molecular Biology and Evolution*

Zhongqi Zhang, Dongping Xu and Xun Gu

**Abstract**

In comparative genomic studies, making phylogenetic trees across species or with gene families has become one of the standard approaches to uncover evolutionary relationships among duplicate genes. From an evolutionary point of view, sequence divergence provides insight about the evolution of new genes after gene duplication which is thought to be related to their function divergence. During the last decade, the development of large-scale gene expression analysis – microarray – enables scientists to simultaneously measure expression profiles of a large set of genes. Since expression profiles measure the transcriptional activity of genes, expression divergence can be used to infer function divergence. In this article, the analysis of gene expression profiles is put into the context of gene evolution and some strategies of distance-based phylogenetic methods are given about how to conduct phylogenetic analysis of microarray data.

**Introduction**

Understanding the evolutionary pattern of gene expression has been greatly facilitated by DNA microarray technology because it can simultaneously monitor expression levels of thousands of genes across many experimental conditions or treatments (e.g. Wagner 2000; Enard et al. 2002; Gu et al. 2002; Gu and Gu 2003; Rifkin et al. 2003). Most evolutionary expression analyses can be classified into two categories. The first one is to explore genome-wide expression evolution between species and the connection to the evolution of species-specific phenotypes from primates to fruitflies (e.g. Enard et al. 2002; Gu and Gu 2003; Rifkin et al. 2003). The second one, on which we are focused here, is to explore the pattern of expression divergence after gene duplication, as well as evolutionary correlations with
other parameters, such as sequence divergence, regulatory motif structure, knock-out phenotypes or protein interactions (e.g. Wagner 2000; Gu et al. 2002; Gu et al. 2002; Gu and Gu 2003; Rifkin et al. 2003). Although significant progress has been made, further study could be hindered by the lack of a clear-cut evolutionary model for expression evolution.

Gu (2004) has developed a stochastic framework for studying the evolution of gene expression under various evolutionary mechanisms. He provided conditions under which the joint density of expression intensity of individual genes (likelihood function) is Gaussian normal; the variance-covariance matrix is determined in terms of the evolutionary model and the known phytogeny. Though the maximum likelihood approach gives more efficient estimates, the numerical algorithm becomes very tedious for large gene families. Moreover, since there are many unknown parameters that must be estimated, the problem of over-parameterization could be serious. From the view of practice, it is desirable to develop some useful but fast approaches to parameter estimation.

Displaying relationships among units in a tree format is visually powerful. Relationships can be easily visualized through the connection of branches, nodes and branch lengths. Reconstructing phylogenetic trees in biological science was first applied to the study of morphological characters to establish evolutionary relationships among species. As the biological sciences advanced into the molecular level, phylogenetic analysis of DNA and protein sequences became an important approach to study evolutionary relationships within a gene family. As large scale gene expression data become available, it is inevitable that it will be incorporated into phylogenetic research. In this paper, we enlist some distance methods, mostly derived from molecular phylogenetic research, which can be adopted into phylogenetic expression studies.

Molecular phylogeny is a well-developed technology for dealing with sequence data. In general, there are three major approaches to reconstruct a phylogenetic tree: distance (matrix) methods, maximum parsimony methods (MP) and maximum likelihood methods (ML)\(^1\). MP and ML methods specifically deal with discrete sequence data; therefore, we will not discuss them here. For the distance methods, although there are many different techniques, all are initiated with the pair-wise distance ($d_{ij}$) matrix. (Later, we will show that,

\(^1\)For details about MP and ML approaches, see “Molecular Evolution and Phylogenetics” by Nei and Kumar, 2000.
after transforming the expression data of a gene family into an expression distance matrix, many ready-to-use methods from molecular phylogenetic study can be easily applied.) One of the simplest distance methods is the “unweighted pair-group method using arithmetic averages (UPGMA)” (Sneath and Sokal 1973). The methodology of UPGMA is the same as “hierarchical clustering with average linkage” in multivariate data analysis. Two taxa with the shortest distance in the distance matrix are grouped together (clustered) and the distances between the cluster and other taxa are computed using a weighted average, reducing the dimension of the distance matrix by one at each step. The iteration continues until there is only one cluster. The second major distance method is called the “least square (LS)” method (Cavalli-Sforza and Edwards 1967; Fitch and Margoliash 1967), which minimizes the residual sum of squares between the observed and estimated pair-wise distances. For example, in figure 1(b), the observed distance between taxa 1 and 5 is \( d_{15} \) in the distance matrix, and the estimated expected distance (\( \hat{E}_{15} \)) between taxa 1 and 5 is 
\[
\hat{E}_1 + \hat{E}_6 + \hat{E}_7 + \hat{E}_8 \text{ where } \hat{E}_i \text{ is estimated using LS.}
\]
This method finds the topology that minimizes \( \sum_{i<j} (d_{ij} - \hat{E}_{ij})^2 \). Although LS can be used to find the topology, the probability of finding the correct topology is rather low compared to some other distance methods (Saitou and Nei 1986; Rzhetsky and Nei 1992). However, the major role of the LS approach is in estimating the branch length of a given topology derived from other methods. The third distance method is the “minimum evolution (ME) method”, which looks for the tree with the smallest sum of branch lengths, i.e. minimizes \( \sum_i E_i \) (Kidd and Sgaramella-Zonta 1971).

The fourth distance method, the “neighbor joining (NJ) method”, is a simplified version of the ME method (Saitou and Nei 1987). The NJ method starts with a star-like tree, where all taxa are connected through just one node. Then, the NJ tree is constructed by successively identifying and joining neighbors (two taxa connected through one node). Note that, none of the above distance methods was developed to identify a rooted phylogeny, i.e. the output of those methods is always an unrooted phylogeny.

---

1 UPGMA can give a rooted phylogeny under the assumption of constant evolutionary rate, which is not always true.
For decades, distance methods have played an indispensable role in molecular evolutionary study, for both tree-making and evolutionary pattern analysis (Nei 1987; Satiou and Nei 1987; Li 1997). When we apply this methodology to the study of gene expression evolution, the critical step is to define an additive expression distance measure between two genes. In this paper we shall address these issues under a general stochastic model for gene expression evolution. Although we focus on expression evolution after gene duplications, it can be extended to orthologous expression analysis across species.

The outline of our study is as follows. We first define the expression distance between a pair of genes based on the E\textsubscript{0} model by Gu (2004) and subsequently define an expression distance matrix. Then, we consider the reconstruction of an expression phylogeny from an expression distance matrix, including dealing with the expression prior at root. Using yeast microarray and gene family data as examples, we demonstrate potential applications of the distance-based methods for phylogenetic analysis of expression data.

Methods

Definition of expression distance

Figure 1. (a) The rooted phylogeny of a five-gene family. (b) The unrooted phylogeny of the same five-gene family in (a), where branches 7 and 8 in the rooted phylogeny are merged as branch 7 in the unrooted case.

\footnote{The additivity of expression distance assures that, if the underlying evolutionary model holds, the pair-wise distance can be appropriately mapped onto the phylogeny by the method of least squares.}
Recently, Gu (2004) set up a statistical framework for phylogenetic analysis of gene expression data, where the variance/covariance matrix of gene expression at the external nodes (\(V_{ij}\)) is modeled as the sum of variations associated with each branch,

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

where \(\rho^2\) is the common variance component of the gene family at root 0; \(E_i\) represents all kinds of variation along branch \(i\), such as the random genetic drift (stochastic variation), the variation associated with directional trend and/or the variation from a dramatic genetic shift after gene duplication (see Gu 2004 for details)\(^1\); \(k \in x_i\) refers to all branches in the lineage from root 0 to \(x_i\) and \(k \in x_i, x_j\) refers to all branches shared by \(x_i\) and \(x_j\) since root 0. In our study, the evolutionary expression variation associated with each branch \(E_i\) is used as a measure of expression branch distance/length. Equ. 1 provides a lineage representation of distance measures, covering the whole lineage from the root to the current taxa. (Later, we show that this representation is especially useful in inferring a rooted expression phylogeny.) We define the pair-wise expression distances \((E_{ij})\) between two taxa as the sum of variations \((E_i)\) along all branches in the path connecting two taxa, i.e.

\[
E_{ij} = \sum_{k \in x_i} E_k + \sum_{l \in x_j} E_l - 2 \sum_{m \in (x_i, x_j)} E_m \\
= (V_{ii} - \rho^2) + (V_{jj} - \rho^2) - 2 (V_{ij} - \rho^2) \\
= V_{ii} + V_{jj} - 2 V_{ij},
\]

where \(k \in x_i, l \in x_j\) and \(m \in (x_i, x_j)\) have the same meaning as in Equ. 1. Note that, although the basic strategy of reconstructing a phylogenetic tree is similar in both sequence data and expression data, there is a fundamental difference about how to deal with the root. In molecular phylogeny, the root is a point within the tree topology, usually identified through the presence of an out group or some existing knowledge about the evolutionary relationship among taxa. Adding a root will not affect the branch lengths of the tree. However, in expression phylogeny, besides being an interior point on the phylogeny, the root itself also

\(^1\) Or see chapter I for a brief review.
contains a certain "distance \( p^2 \)", i.e. expression prior, slightly changing the branch lengths within the tree. Since identifying a root is usually not an easy task, we will discuss the unrooted and rooted cases separately.

**Phylogenetic inference**

With the definition of expression distance available, many approaches already available in molecular phylogenetic study – distance methods – can be directly applied to the expression phylogenetic analysis.

For expression phylogenetic analysis, there are two basic strategies. First, considering the potential noise associated with expression data and the rather well-developed techniques in molecular phylogeny, the topology of the phylogeny is reconstructed based on DNA/protein sequence data and the expression distance is mapped onto the topology using the least-squares (LS) approach, called sequence-guided phylogenetic expression analysis (sgE). In this way, the accuracy of the evolutionary relationship among duplicate genes is preserved. Second, the expression phylogeny can be inferred solely from the expression distance matrix, called an E tree, without the aid of molecular information.

(1) Sequence-guided expression tree (sgE tree)

For the sequence-guided phylogeny, least square (LS) method is the standard approach to replace the molecular branch lengths by the expression distances. The original LS method for a given topology is to solve a series of linear equations, usually with more equations than unknown parameters.

We first discuss the expression mapping of an unrooted phylogeny. Using an unrooted 5-gene phylogeny as an example (figure 1b), the pair-wise distance between two taxa \( i \) and \( j \) can be written as the sum of branches linking the two genes:

\[
\begin{align*}
    d_{12} &= E_1 + E_2 + e_{12} \\
    d_{13} &= E_1 + E_3 + E_6 + e_{13} \\
    d_{14} &= E_1 + E_4 + E_6 + E_7 + e_{14} \\
    d_{15} &= E_1 + E_5 + E_6 + E_7 + e_{15} \\
    d_{23} &= E_2 + E_3 + E_6 + e_{23}
\end{align*}
\]
\[ d_{24} = E_2 + E_4 + E_6 + E_7 + e_{24} \]
\[ d_{25} = E_2 + E_5 + E_6 + E_7 + e_{25} \]
\[ d_{34} = E_3 + E_4 + E_7 + e_{34} \]
\[ d_{35} = E_3 + E_5 + E_7 + e_{35} \]
\[ d_{45} = E_4 + E_5 + E_7 + e_{45} \]

where \( d_y = E_y + e_y = \hat{V}_y + \hat{V}_y - 2\hat{V}_y \), \( E_y \) is as in Equ. 2, \( e_y \) is the measurement error associated with \( d_y \) and \( \hat{V} \) is the variance/covariance matrix for gene expression at external nodes. Equation (3) can be written in matrix form as

\[
\begin{bmatrix}
  d_{12} \\ d_{13} \\ d_{14} \\ d_{15} \\ d_{23} \\ d_{24} \\ d_{25} \\ d_{34} \\ d_{35} \\ d_{45}
\end{bmatrix} =
\begin{bmatrix}
  1 & 1 & 0 & 0 & 0 & 0 & 0 \\
  1 & 0 & 1 & 0 & 0 & 1 & 0 \\
  1 & 0 & 0 & 1 & 0 & 1 & 1 \\
  1 & 0 & 0 & 0 & 1 & 1 & 1 \\
  0 & 1 & 1 & 0 & 0 & 1 & 0 \\
  0 & 1 & 0 & 1 & 0 & 1 & 1 \\
  0 & 1 & 0 & 0 & 1 & 1 & 1 \\
  0 & 0 & 1 & 1 & 0 & 0 & 1 \\
  0 & 0 & 1 & 0 & 1 & 0 & 1 \\
  0 & 0 & 0 & 1 & 1 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
  E_1 \\
  E_2 \\
  E_3 \\
  E_4 \\
  E_5 \\
  E_6 \\
  E_7 \\
\end{bmatrix}
+ e_d
\]

(4)

or \( \mathbf{d} = X \hat{E} + e_d \).

The LS estimator of \( \hat{E} \) is

\[ \hat{E} = (X^T X)^{-1} X^T \mathbf{d}. \]

When the number of genes in a gene family is large, computing the inverse of \( (X^T X) \) will take a large amount of computational time and will be sensitive to round-off error. Rzhetsky and Nei (1993) proposed a simple and fast algorithm to compute branch length without the use of matrix algebra (also see Gascuel 1997, Bryant and Waddell 1998). The basic strategy is to reduce the topology to a four-taxa phylogeny and compute the interior branch length analytically; for an exterior branch, the tree topology is reduced to a three-taxa phylogeny to compute the exterior branch length analytically (see Appendix for a brief summary). There is also an approximate algorithm, called “Fitch and Margoliash’s method”, which speeds up all

---

1 If the measurement errors are i.i.d. with variance \( \sigma^2 \), \( e_y = 2\sigma^2 \).
branch length computations by reducing the topology into a three-taxa tree (see Fitch 1967 for detail). In most cases, the Fitch and Margoliash’s method gives similar branch length estimates to the LS approach.

Next, we turn our attention to the rooted phylogeny. In most tree making methods, the root of a phylogeny is hard to identify unless the assumption of constant evolutionary rates among branches is imposed, which is not always true. Therefore, the position of the root is inferred based on knowledge other than sequence information, typically using out groups which are known to split from genes of interest in a more ancient time. After a rooted phylogeny has been constructed, mapping the expression distances onto the topology is fairly easy, similar to equation 4. The lineage representation (Equ. 1) of expression distances is used in LS estimation instead of pair-wise distances. For the rooted 5-gene phylogeny in figure 1a,

\[
\begin{bmatrix}
\hat{V}_{11} \\
\hat{V}_{12} \\
\hat{V}_{13} \\
\hat{V}_{14} \\
\hat{V}_{15} \\
\hat{V}_{22} \\
\hat{V}_{23} \\
\hat{V}_{24} \\
\hat{V}_{25} \\
\hat{V}_{33} \\
\hat{V}_{34} \\
\hat{V}_{35} \\
\hat{V}_{44} \\
\hat{V}_{45} \\
\hat{V}_{55}
\end{bmatrix}
= \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
\end{bmatrix}
+ \begin{bmatrix}
\rho^2 \\
E_1 \\
E_2 \\
E_3 \\
E_4 \\
E_5 \\
E_6 \\
E_7 \\
E_8 \\
E_9 \\
E_{10} \\
E_{11} \\
E_{12} \\
E_{13} \\
E_{14} \\
E_{15} \\
E_{16} \\
E_{17} \\
E_{18} \\
E_{19} \\
E_{20} \\
E_{21} \\
E_{22} \\
E_{23} \\
E_{24} \\
E_{25} \\
E_{26} \\
E_{27} \\
E_{28} \\
E_{29} \\
E_{30} \\
E_{31} \\
E_{32} \\
E_{33} \\
E_{34} \\
E_{35} \\
E_{36} \\
E_{37} \\
E_{38} \\
E_{39} \\
E_{40} \\
E_{41} \\
E_{42} \\
E_{43} \\
E_{44} \\
E_{45} \\
E_{46} \\
E_{47} \\
E_{48} \\
E_{49} \\
E_{50} \\
E_{51} \\
E_{52} \\
E_{53} \\
E_{54} \\
E_{55}
\end{bmatrix}
\]

or \( V = WE + \varepsilon \).
where \( \hat{\epsilon}_g = V_{ij} + \epsilon_g \) and \( \epsilon_g \) is the measurement error\(^1\). The LS estimate of \( \beta^* \),
\[
\hat{\beta}^* = (W^T W)^{-1} W^T \hat{\epsilon}.
\]
There is also a fast algorithm to compute the estimates of the expression branch length for a rooted phylogeny, including the root prior variance (\( \hat{\rho}^2 \)) (see Appendix).

(2) Expression tree (E tree) solely based on expression information

When both the sequence information and the expression information are available, the preferred approach of reconstructing an expression tree is to infer the topology using the sequence information first then map the expression distance onto that topology (sqE tree) as described in the previous section. For cases when no sequence information is available, it is also possible to reconstruct the expression tree only based on the expression distance matrix (E matrix). Given the E matrix, many ready-to-use distance methods in molecular phylogenetic analysis can be directly adopted, such as UPGMA, LS, ME and NJ (see corresponding molecular phylogenetic analysis – distance methods – for details). Note that it is usually difficult to infer the root of an E tree based on the expression data alone\(^2\).

(3) Evaluating the reliability of the expression tree

For an expression tree (either sqE tree or E tree), there is always some uncertainty associated with both the tree topology and the branch length estimation. Felsenstein (1985) proposed a bootstrap method which is used in molecular phylogenetic analysis to address uncertainty. Here, we use microarray gene expression data as an example to demonstrate the use of bootstrapping in evaluating estimated gene expression trees. The input microarray data is a collection of expression measurements of \( n \) genes under \( G \) conditions. The \( n \) expression measurements under the same condition are treated as one unit. Assuming independence across \( G \) conditions (units) and that the conditions are a random sample from some population of possible conditions, a bootstrap analysis is conducted with \( G \) units selected by

---

\(^1\) Similarly as in unrooted case, if measurement error is i.i.d. with variance \( \sigma^2 \), \( \epsilon_g = \sigma^2 \) if \( i = j \); 0 otherwise.

\(^2\) The E tree is constructed using distance methods, which do not infer the root of the tree unless under the assumption of a constant rate of expression evolution. (See Nei and Kumar 2000 for details about dealing with the root using the distance methods.)
random sampling with replacement in each of \( B \) independent bootstrap samples. A distance matrix is computed for each bootstrap sample, and an expression tree (either an sqE tree or an E tree depending on the original analysis) is constructed. When evaluating the reliability of the expression tree topology (E tree)\(^1\), the percentage of bootstrap topologies in agreement with the original E tree topology is used as an indicator of the topology reliability. For example, if 90% of bootstrap samples give the same topology as the original data, we have a 90% confidence that the topology is reliable. Meanwhile, the sampling variance of any expression distance \( E_{ij} \) or any expression branch length \( E_k \), can be calculated by the bootstrapping procedure.

**Enolase (Eno) Gene Family of *Saccharomyces cerevisiae*: an Example**

The enolase (Eno) gene family was chosen as an example to demonstrate the application of phylogenetic expression analysis. Enolase is an enzyme catalyzing the dehydration of 2-phosphoglycerate (2PG) to phosphoenolpyruvate (PEP) in glycolysis, also the reverse reaction in Gluconeogenesis. Based on the COG (cluster of orthologous groups) gene family profiles of NCBI, the Eno gene family is composed of 50 genes from 25 species (http://www.ncbi.nlm.nih.gov/COG/afn/COG0148.aln). Among those 50 genes, 5 genes are yeast genes: YGR254w (enolase I), YHR174w (enolase II), YMR323w (hypothetical ORF), YOR393w (unknown function) and YPL281c (unknown function). Figure 2 shows the molecular phylogeny of these five genes based on their protein sequence data. The NJ tree and the UPGMA tree gave the same topology and similar branch lengths. To match the notation in figure 1, we have YOR393w=Gene1, YPL281c=Gene2, YMR323w=Gene3, YGR254w=Gene4 and YHR174w=Gene5.

---

\(^1\) Since the topology of an sqE tree is based on molecular phylogeny that is treated as given, we do not evaluate the reliability of the topology of an sqE tree using the expression data.
To locate the root position of Eno phylogeny, we reconstructed the molecular phylogeny of the whole COG Eno gene family as shown in figure 3. It is obvious that the root is somewhere in the longest interior branch.
Figure 3. Neighbor-joining (NJ) tree of enolase gene family (COG0148). The highlighted branches are the five-gene phylogeny of Saccharomyces cerevisiae.

Microarray data were downloaded from online data sources (see references). Totally 5 x 230 data points are available for this 5-yeast-gene family (Table 1). An estimate of the variance-covariance matrix (Table 2) was computed as,
\[ \hat{V}_{ij} = \frac{1}{G-1} \sum_{g=1}^{G} (x_{ig} - \bar{x}_i)(x_{jg} - \bar{x}_j) \]

where \( G = 230 \), \( x_{ig} \) is the expression measure for gene \( i \) under experimental condition \( g \) and \( \bar{x}_i \) (\( \bar{x}_j \)) is the mean expression of gene \( i \) (\( j \)) across all \( G \) conditions. Table 3 is the pair-wise distance matrix: \( d_{ij} = \hat{V}_{ii} + \hat{V}_{jj} - 2 \cdot \hat{V}_{ij} \).

Using LS estimation for a rooted tree (Equ. 5, Table 2) and unrooted tree (Equ. 4, Table 3), we constructed both the rooted sqE tree and the unrooted sqE tree (figure 4). In both phylogenies, nodes 6 and 7 merge into just one node, i.e. the estimated expression branch 6 (\( \hat{E}_6 \)) is about zero. Table 4 provides some detailed information about expression branch length under several tree making strategies. Bootstrap was used to evaluate the reliability of the expression tree and the sampling variance of estimated expression distances. For example, using (unrooted) sqE tree making method, we observed that \( \hat{E}_5 \) is about 3 times longer than \( \hat{E}_4 \) (Table 4), which indicates that the expression variation along the YHR174w (enolase II) lineage is higher than in the YGR254w (enolase I) lineage; however, the 95% bootstrap confidence interval for \( \hat{E}_5 - \hat{E}_4 \) is (-0.04, 0.50), showing that this difference is not significant at 0.05 confidence level. In other words, the expression evolutionary rates after gene duplication between YHR174w and YGR254w are not statistically distinguishable.
Table 1. Part of the microarray data of 5 yeast genes in Eno gene family.

<table>
<thead>
<tr>
<th>UID</th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
<th>G5</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>YGR254W</td>
<td>0.0144</td>
<td>-0.0893</td>
<td>-0.0589</td>
<td>0.0286</td>
<td>0.3785</td>
<td>...</td>
</tr>
<tr>
<td>YHR174W</td>
<td>-0.0893</td>
<td>-0.2345</td>
<td>-0.2515</td>
<td>-0.0589</td>
<td>-0.2345</td>
<td>...</td>
</tr>
<tr>
<td>YMR323W</td>
<td>0.1506</td>
<td>0.0286</td>
<td>0.1375</td>
<td>0.4114</td>
<td>0.3896</td>
<td>...</td>
</tr>
<tr>
<td>YOR393W</td>
<td>0</td>
<td>-0.0291</td>
<td>-0.2689</td>
<td>0.2750</td>
<td>0.5261</td>
<td>...</td>
</tr>
<tr>
<td>YPL281C</td>
<td>0.0841</td>
<td>-0.0145</td>
<td>-0.1047</td>
<td>0.0566</td>
<td>0.2987</td>
<td>...</td>
</tr>
</tbody>
</table>

The data are measured by log2(fold change of gene expression level between treatment and control).

Table 2. The estimated variance-covariance matrix (\( \hat{V} \)) of 5 yeast genes in Eno gene family.

<table>
<thead>
<tr>
<th>( \hat{V}_{ij} )</th>
<th>YOR393w</th>
<th>YPL281c</th>
<th>YMR323w</th>
<th>YGR254w</th>
<th>YHR174w</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOR393w</td>
<td>0.584034</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YPL281c</td>
<td>0.276619</td>
<td>0.560165</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YMR323w</td>
<td>0.291283</td>
<td>0.379585</td>
<td>0.457886</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YGR254w</td>
<td>-0.17858</td>
<td>-0.14011</td>
<td>-0.10562</td>
<td>0.624339</td>
<td></td>
</tr>
<tr>
<td>YHR174w</td>
<td>-0.15867</td>
<td>-0.07188</td>
<td>-0.04789</td>
<td>0.568459</td>
<td>0.949816</td>
</tr>
</tbody>
</table>

Table 3. The pair-wise distance matrix of 5 yeast genes in Eno gene family

<table>
<thead>
<tr>
<th>( d_{ij} )</th>
<th>YOR393w</th>
<th>YPL281c</th>
<th>YMR323w</th>
<th>YGR254w</th>
<th>YHR174w</th>
</tr>
</thead>
<tbody>
<tr>
<td>YOR393w</td>
<td></td>
<td>0.5910</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YPL281c</td>
<td>0.4594</td>
<td>0.2589</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YMR323w</td>
<td>1.5655</td>
<td>1.4647</td>
<td>1.2935</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YGR254w</td>
<td>1.8512</td>
<td>1.6537</td>
<td>1.5035</td>
<td>0.4372</td>
<td></td>
</tr>
<tr>
<td>YHR174w</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4 (a). Summary of branch length in sgE tree-making methods.

<table>
<thead>
<tr>
<th>( \rho^2 )</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
<th>E6</th>
<th>E7</th>
<th>E8</th>
</tr>
</thead>
<tbody>
<tr>
<td>rooted sgE</td>
<td>0(^\d)</td>
<td>0.31</td>
<td>0.28</td>
<td>0.12</td>
<td>0.06</td>
<td>0.38</td>
<td>0(^\d)</td>
<td>0.45</td>
</tr>
<tr>
<td>unrooted sgE</td>
<td>NA</td>
<td>0.38</td>
<td>0.21</td>
<td>0.06</td>
<td>0.11</td>
<td>0.32</td>
<td>0(^\d)</td>
<td>1.12</td>
</tr>
</tbody>
</table>

\(^\d\)Negative LS estimations are set to 0.

Table 4 (b). Summary of branch length in different E tree-making methods.

<table>
<thead>
<tr>
<th>( \rho^2 )</th>
<th>E1</th>
<th>E2</th>
<th>E3</th>
<th>E4</th>
<th>E5</th>
<th>E6</th>
<th>E7</th>
<th>E8</th>
</tr>
</thead>
<tbody>
<tr>
<td>E (NJ)(^\d)</td>
<td>NA</td>
<td>0.06</td>
<td>0.20</td>
<td>0.38</td>
<td>0.11</td>
<td>0.33</td>
<td>0.02</td>
<td>1.11</td>
</tr>
<tr>
<td>E (UPGMA)(^\d)</td>
<td>NA</td>
<td>0.13</td>
<td>0.13</td>
<td>0.26</td>
<td>0.22</td>
<td>0.22</td>
<td>0.13</td>
<td>1.07</td>
</tr>
</tbody>
</table>

\(^\d\)The tree topology of sgE tree is based on figure 1 and figure 2, where gene1=YOR393w, gene2=YPL281c, gene3=YMR323w, gene4=YGR254w and gene5=YHR174w.

\(^\d\)The tree topology of E tree is based on figure 1(b) and figure 5, where gene1= YMR323w, gene2=YPL281c, gene3= YOR393w, gene4=YGR254w and gene5=YHR174w, which is different from sgE tree topology.
Figure 4. (a) The rooted sequence-guided expression tree of five yeast genes in Eno gene family. (b) The unrooted sequence-guided expression tree. In both cases, branch 6 (E_6 in figure 1) is nearly zero and nodes 6 and 7 merge together.

When the sequence information is not available, we can reconstruct the expression phylogeny using the distance methods described previously. However, caution needs to be exercised because there is some discrepancy between sequence tree topology and expression (E) tree topology. Suppose we do not have the sequence information for Eno gene family. Figure 5 shows the expression tree (E tree) based on the expression distance matrix (Table 3) using NJ and UPGMA. Both E trees have the same topology, but they are slightly different from sgE trees in figure 4, especially when looking at the topology associated with YOR393w, YPL281c and YMR323w. (However, when looking at the confidence interval for those branches, we see consistent results that nodes 6 and 7 are merged together.) In the E tree, we labeled YMR323w=Gene1, YPL281c=Gene2, YOR393w=Gene3,
YGR254w=Gene4 and YHR174w=Gene5 (figure 5); therefore, the branch annotation in Table 4 is slightly different between sgE trees and E trees.

![Expression tree diagram](image)

Figure 5. Expression tree using the pair-wise expression distance matrix. (a) Neighbor-joining tree (NJ). (b) UPGAM tree. The root in the UPGMA tree is located based on the assumption of constant rate of expression evolution.

**Discussion**

So far, we have listed several distance-based methods useful in phylogenetic expression analysis. These methods, as in the molecular phylogenetic analysis, could provide the starting point for exploring expression evolution relationships among duplicate genes. We have recognized that the expression distance involves several assumptions that need to be examined carefully. By default, we assume that \( \log_2(\text{fold change of gene expression}) \) follows a Brownian-based process during evolution\(^1\). Further investigation is needed to test the robustness of normal assumption when other data normalization/transformation procedures are adopted (Quackenbush 2001).

**Effect of experimental noise**

One assumption in the \( E_0 \) model is that the measurement error is small compared to other variance components in the model (Gu 2004)\(^2\). In practice, this assumption may or may not be true. We show that, under i.i.d. error assumption, LS estimates of interior branch

---

\(^1\) Here, we assume the transformed measurements to be normally distributed.

\(^2\) Also see chapter I for some comments.
lengths are robust against the presence of measurement error; while the estimates of exterior branches are not.

Assuming that experimental errors\(^1\) are i.i.d. with variance \(\sigma^2_e\), the estimates of the elements of the variance-covariance matrix have the form

\[
E(V_{ij}) = \begin{cases} 
V_{ij} + \sigma^2_e & \text{if } i = j \\
V_{ij} & \text{if } i \neq j
\end{cases}
\]  

(7)

And the estimates of the pair-wise distances

\[
E(d_{ij}) = E_{ij} + 2\sigma^2_e.
\]

(8)

Both \(V_{ij}\) and \(E_{ij}\) tend to be over-estimated if no correction is made for measurement error. For the unrooted expression phylogenetic analysis using LS, it can be shown that the interior branch length is not affected by such over-estimation (see Appendix a1). But for exterior branches, they are over-estimated with an error term \(\sigma^2_e\) (see Appendix a2). The same conclusions hold for rooted expression phylogenetic analysis using LS. In Appendix (a3) and (a4), the estimates of the expression distance from the root to an exterior node (\(\hat{D}_i\)) are shown to be over-estimated with error \(\sigma^2_e\); while the estimates are unbiased for interior nodes (including \(\rho^2\) at root). Since the interior branch length is the difference between the expression distances of two interior nodes (Appendix a6), it is unbiased. Since the expression distance of an exterior branch is a difference between the expression distance (\(D\)) of an exterior node and the expression distance of an interior node, and it is over-estimated by \(\sigma^2_e\).

Therefore, in both rooted and unrooted phylogeny, expectations of estimated branch lengths are

\[
E(\hat{E}_i) = \begin{cases} 
E_i + \sigma^2_e & \text{if } \text{gene } i \text{ is an exterior gene} \\
E_i & \text{otherwise}
\end{cases}
\]

(9)

When the variance of the measurement error can be estimated (e.g. the original microarray data have replicates), corresponding correction can be made with exterior branch length.

**Relationships with molecular phylogenetics**

\(^1\) Obviously, the variance of experimental errors does not depend on the tree topology.
Throughout our gene expression phylogenetic analysis, we adopted some basic methodology from molecular phylogenetics. Except for the definition of expression distance and the issue of dealing with expression distance at a root, almost every step in expression analysis has a counterpart in molecular phylogenetics. In Table 5, we summarize some differences and similarities between gene expression phylogenetic analysis and molecular phylogenetic analysis.

### Table 5. Comparison between expression phylogenetic and molecular phylogenetic analyses.

<table>
<thead>
<tr>
<th></th>
<th>Expression Phylogenetics</th>
<th>Molecular Phylogenetics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pair-wise distance</strong></td>
<td>$V_{ii} + V_{jj} - 2V_{ij}$ based on $E_0$ (Brownian-based) model</td>
<td>e.g. $-\ln (1-p)$ based on the Poisson distribution*</td>
</tr>
<tr>
<td><strong>between taxa</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Topology</strong></td>
<td>Either from molecular phylogeny or using the distance methods</td>
<td>Distance methods, parsimony methods and likelihood methods</td>
</tr>
<tr>
<td><strong>Branch length</strong></td>
<td>Least Squares</td>
<td>Least Squares(^\dagger)</td>
</tr>
<tr>
<td><strong>estimation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Distance associated</strong></td>
<td>Rooted expression phylogenetics</td>
<td>N.A.</td>
</tr>
<tr>
<td><strong>with root</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* $p$ is the proportion of different amino acids between two sequences (Grishin 1995)

\(^\dagger\) LS method is the standard method for estimating the branch lengths of a tree in molecular phylogenetics

We have shown that gene expression analysis alone can reconstruct the expression phylogenetic relationships among duplicate genes, i.e. E tree. The question is if expression phylogenetic analysis gives the same topology as in molecular phylogenetic analysis. Our analysis showed that it is not uncommon to have discrepancy between sequence topology and expression topology which, we think, could be well explained by the decoupling between sequence divergence and expression divergence. For example, we did both molecular and expression phylogenetic analyses with all four-gene yeast families in the COG. Among the 38 COG gene families, only 17 families (45%) have the same topology in both analyses. In the future, it will be interesting to address such decoupling on a large scale, e.g. all gene
families in yeast, and look for how the interaction between these two divergence processes shapes the evolution paths of the duplicate genes.

References
Grishin, N.V. (1995) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J. Mol. Evol. 41, 675-679

On-line data source:
http://www.genome.stanford.edu/yeast_stress
http://cellcycle-www.stanford.edu


Appendix

Fast algorithm for evaluation of branch length estimates for an unrooted phylogeny

In 1993, Rzhetsky and Nei proposed a fast algorithm to evaluate the LS estimates of branch lengths without using matrix inversion. For a given topology with more than 4 taxa, when computing an interior branch length, the phylogeny is reduced to a standard 4-gene topology as shown above in figure A1(b); or reduced to a standard 3-gene topology as shown in A1(c) when computing an exterior branch length. For example, if we want to compute the branch length of the interior branch $E_\ell$ in figure A1(a), we have $A=\{1\}, B=\{2\}, C=\{3\}$ and $D=\{4,5\}$ as in figure A1(b); if we want to compute the exterior branch $E_\ell$ (figure A1(a)), we have $A=\{1\}, B=\{2\}$ and $C=\{3,4,5\}$ as in figure A1(c).

The LS estimate of an interior branch length (figure A1(b)) is given by

$$E = \frac{1}{2} \left\{ \gamma \left[ \frac{d_{AC}}{m_A m_C} + \frac{d_{BD}}{m_B m_D} \right] + (1 - \gamma) \left[ \frac{d_{BC}}{m_B m_C} + \frac{d_{AD}}{m_A m_D} \right] - \frac{d_{AB}}{m_A m_B} - \frac{d_{CD}}{m_C m_D} \right\}$$

(a)

$$\gamma = \frac{m_B m_C + m_A m_D}{(m_A + m_B)(m_C + m_D)}$$

where $m_I$ is the number of taxa in cluster $I$ ($I = A, B, C$ or $D$) and $d_{IJ}$ is the sum of all pairwise distances between cluster $I$ and cluster $J$.

On the other hand, the LS estimate of an exterior branch length (figure A1(c)) is given by

Figure A1. LS estimation of branch lengths in an unrooted phylogeny. In (b), A, B, C and D can be a cluster of genes; while in (c), B and C can be a cluster of genes.
\[ E = \frac{1}{2} \left( \frac{d_{AB}}{m_B} + \frac{d_{AC}}{m_C} - \frac{d_{BC}}{m_Bm_C} \right) \]  

where \( m_i \) and \( d_{ij} \) are the same as above. (Note that \( A \) in figure A.1(c) is always a cluster of only one taxa.)

**Fast algorithm in computing the expression branch length of a rooted phylogeny**

Before providing details about the algorithm, we first clarify some notation. First, the nearest ancestral node \((N_a)\) of a branch is defined as the first ancient node upstream of the query node. For example, in figure 1(a), the \( N_a \) of node 7 is the root; the \( N_a \) for node 4 is node 8; and there is simply no \( N_a \) for the root. Similarly, we define the nearest progeny node \((N_p)\) of a branch as the first node downstream. Second, the offspring cluster of a node refers to the two clusters \((C_1 \text{ and } C_2)\) of exterior taxa after the bifurcate duplication event at that node. As in figure 1(a), the offspring clusters of node 0 are \((1,2,3)\) and \((4,5)\); the offspring clusters of node 7 are \((1,2)\) and \((3)\); and there are no offspring clusters for an exterior node. For an interior node \( i \), the expression distance \((D_i)\) from the root up to that node is defined as the average of all pair-wise covariances \((V_{kl})\) between its two offspring clusters \( C_1 \) and \( C_2 \), i.e.

\[
D_i = \frac{\sum_{k \in C_1} \sum_{l \in C_2} V_{kl}}{m_{C_1}m_{C_2}}
\]

where \( m_{C_1} \) and \( m_{C_2} \) are the numbers of taxa in the offspring clusters. By contrast, the expression distance up to an exterior node \( i \) is given by its expression variance, i.e.

\[
D_i = V_{ii}
\]

Using the phylogeny in figure 1(a) as an example, we have

\[
\begin{align*}
D_1 &= V_{11}, \quad D_2 = V_{22}, \quad D_3 = V_{33}, \quad D_4 = V_{44}, \quad D_5 = V_{55} \\
\rho^2 &= D_0 = \frac{V_{14} + V_{15} + V_{24} + V_{25} + V_{34} + V_{35}}{3 \times 2} \\
D_6 &= V_{12} \\
D_7 &= \frac{V_{13} + V_{23}}{2 \times 1} \\
D_8 &= V_{45}
\end{align*}
\]
The LS estimate of expression branch length \( E_i \) is simply the difference in estimated distances between its nearest progeny node \( (N_p) \) and nearest ancestral node \( (N_a) \),

\[
E_i = D_{N_p} - D_{N_a}
\]  

(a6)

where \( D_{N_a} \) of the root is 0. For the 5-gene family in figure 1(a), we have

\[
\begin{align*}
E_1 &= V_{11} - V_{12} \\
E_2 &= V_{22} - V_{12} \\
E_3 &= V_{33} - \frac{V_{13} + V_{23}}{2} \\
E_4 &= V_{44} - V_{45} \\
E_5 &= V_{55} - V_{45} \\
E_6 &= V_{12} - \frac{V_{13} + V_{23}}{2} \\
E_7 &= \frac{V_{13} + V_{23}}{2} - \frac{V_{14} + V_{15} + V_{24} + V_{25} + V_{34} + V_{35}}{6} \\
E_8 &= V_{45} - \frac{V_{14} + V_{15} + V_{24} + V_{25} + V_{34} + V_{35}}{6} \\
\rho^2 &= \frac{V_{14} + V_{15} + V_{24} + V_{25} + V_{34} + V_{35}}{6}.
\end{align*}
\]  

(a7)
CHAPTER IV: HOW MUCH EXPRESSION DIVERGENCE AFTER YEAST GENE DUPLICATION COULD BE EXPLAINED BY REGULATORY MOTIF EVOLUTION?

A paper accepted by Trends in Genetics

Zhongqi Zhang, Jianying Gu and Xun Gu

Abstract

We utilized yeast genome sequences of gene families, microarray profiles and regulatory motif data to test the current wisdom that there is a strong correlation between regulatory motif structure and gene expression profile. Our results suggest that duplicate genes tend to be co-expressed but the correlation between motif content and expression similarity is generally poor, only about 2-3% of expression variation can be explained by the motif divergence. Our observations suggests that, in addition to the (cis)-regulatory motif structure in the upstream region of the gene, multiple trans acting factors in the gene network may significantly influence the pattern of gene expression.

Introduction

Predicting the transcriptional regulation network from genomic data is a major challenge [1-5]. Many computational methods are based on the expression-motif-conservation hypothesis that co-expressed genes are likely to share regulatory motifs that are conserved during evolution [6-14]. Several studies have tested this idea [15,16]. For instance, Yu et al. [15] found that (evolutionarily unrelated) yeast gene pairs targeted by shared transcription factors (TF-target gene pair for short), are co-expressed more frequently than expected by chance. One would expect a similar co-expression between duplicated genes, as they usually retain similar function or regulatory pattern.

We used yeast (Saccharomyces cerevisiae) gene families to test this prediction, by examining 202 yeast two-member gene families (BOX 1). Using methods modified from Yu et al. [15], we have shown that 7.0% of duplicate genes pairs are co-expressed (BOX 1),
which is about 7 times greater than random expectation ($P$-value $= 1.9 \times 10^{-6}$). This can be compared to Yu’s finding [15] that “overall, 3.3% TF-target gene pairs are co-expressed, which is 4-times greater than random expectation.” Our analysis indicates that duplicate gene pairs may have even higher co-expression levels compared to TF-target gene pairs. Thus, it is worthwhile exploring a new motif-search strategy by combining co-expressed cluster analysis and gene family phylogeny to increase the signal-noise ratio.

According to the expression-motif-conservation hypothesis, a positive correlation is expected between the fraction of shared motifs and the gene expression similarity for duplicate gene pairs, because both are negatively correlated with the age of duplication event [16, 17]. To minimize the error caused by falsely predicted TF binding sites, we searched the upstream regions of duplicate genes for 50 “known” regulatory motifs compiled by Kellis [14] that are strongly supported by the experiments (BOX1). The mean number of motifs appearing in the upstream region is 8.3 ($\pm 2.4$) per gene, while the mean number of shared motifs between duplicate genes is 1.8 ($\pm 1.2$) per pair. In addition, 183 gene families (91%) have at least one motif with more than one copy; on average, each regulatory motif is present in 1.3 copies in the upstream region. Meanwhile, the expression similarity between two duplicates for each gene family is computed based on 276 microarray data points under various experimental conditions [18].
Fig. 1 (a) A weak but significant (positive) correlation is observed between the fraction of shared motifs and gene expression correlation, showing that the two evolutionary processes are only weakly coupled. (b) A similar pattern between the fraction of “paralogous motifs” and expression correlation. (c) A significant (negative) correlation between the gene expression correlation and the relative evolutionary time. (d) A significant (negative) correlation between the fraction of shared motifs and the relative evolutionary time.
Table 1. Correlations between motif conservation, gene expression similarity and evolutionary time. Gene expression correlation (Corr) between duplicate genes is used as a measurement of expression similarity, while relative evolutionary time (Age) and/or protein distance (Dist) are used to approximate the evolutionary time. The following notation is used in the table: Pearson correlation coefficient (P-value) / Spearman rank correlation (P-value).

<table>
<thead>
<tr>
<th></th>
<th>Corr</th>
<th>Age</th>
<th>Dist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fraction of</td>
<td>0.15 (0.03)</td>
<td>-0.22 (0.001)</td>
<td>-0.20 (0.004)</td>
</tr>
<tr>
<td>shared motifs</td>
<td>/ 0.14 (0.05)</td>
<td>/ -0.19 (0.008)</td>
<td>/ -0.18 (0.009)</td>
</tr>
<tr>
<td>Shared motifs</td>
<td>0.12 (0.09)</td>
<td>-0.15 (0.03)</td>
<td>-0.17 (0.02)</td>
</tr>
<tr>
<td></td>
<td>/ 0.11 (0.10)</td>
<td>/ -0.11 (0.11)</td>
<td>/ -0.15 (0.03)</td>
</tr>
<tr>
<td>Fraction of</td>
<td>0.17 (0.05)</td>
<td>-0.20 (0.004)</td>
<td>-0.19 (0.007)</td>
</tr>
<tr>
<td>shared motifs</td>
<td>/ 0.15 (0.03)</td>
<td>/ -0.18 (0.009)</td>
<td>/ -0.18 (0.009)</td>
</tr>
<tr>
<td>Shared motifs</td>
<td>0.13 (0.06)</td>
<td>-0.14 (0.05)</td>
<td>-0.17 (0.01)</td>
</tr>
<tr>
<td>(weighted)</td>
<td>/ 0.14 (0.06)</td>
<td>/ -0.12 (0.09)</td>
<td>/ -0.17 (0.015)</td>
</tr>
<tr>
<td>Fraction of</td>
<td>0.18 (0.013)</td>
<td>-0.19 (0.006)</td>
<td>-0.20 (0.005)</td>
</tr>
<tr>
<td>“paralogous”</td>
<td>/ 0.16 (0.026)</td>
<td>/ -0.19 (0.006)</td>
<td>/ -0.19 (0.007)</td>
</tr>
<tr>
<td>motifs</td>
<td>0.17 (0.018)</td>
<td>-0.16 (0.02)</td>
<td>-0.18 (0.01)</td>
</tr>
<tr>
<td>“paralogous”</td>
<td>/ 0.15 (0.032)</td>
<td>/ -0.16 (0.02)</td>
<td>/ -0.16 (0.02)</td>
</tr>
<tr>
<td>motifs</td>
<td>0.13 (0.06)</td>
<td>-0.14 (0.05)</td>
<td>-0.17 (0.01)</td>
</tr>
<tr>
<td>Corr</td>
<td>-0.42 (&lt;0.0001)</td>
<td>-0.41 (&lt;0.0001)</td>
<td>-0.43 (&lt;0.0001)</td>
</tr>
<tr>
<td></td>
<td>/ -0.44 (&lt;0.0001)</td>
<td>/ -0.43 (&lt;0.0001)</td>
<td>/ -0.43 (&lt;0.0001)</td>
</tr>
</tbody>
</table>

Fig. 1a shows a weak positive correlation between the fraction of shared motifs and expression similarity (Pearson correlation: \( R = 0.15 \), P-value ~ 0.032; Spearman rank correlation: \( \rho = 0.14 \), P-value ~ 0.048). The regulatory motif model may be too simplistic, e.g., one may consider whether the order and proximity of these motifs is important [19, 20]. Also, a small (but certain) portion of these short motifs could be the result of chance. In spite of its complexity, we have tried several approaches to see whether the motif-expression correlation can be improved.
Previous studies showed that multi-copied regulatory motifs are more likely to be active \[4, 21\]. We have considered this effect by using the number of copies as a weighting factor (BOX1), resulting in a slightly higher statistical significance: Pearson correlation: $R = 0.17$, $P$-value $\sim 0.015$; or Spearman rank: $\rho = 0.15$, $P$-value $\sim 0.032$.

Moreover, the motif-expression correlation becomes more meaningful for “paralogous” motifs between duplicate pairs if the relative position of motifs in the regulatory region is taken into account. Tentatively, we define two similar regulatory motifs located in each duplicate gene as “paralogous” if (a) their positions (relative to the transcription site) are close to each other and (b) they are in the conserved region of yeast evolution (BOX 1). In total, 93 shared paralogous motifs were detected within 72 gene families. The correlation (Fig. 1b) between the fraction of shared paralogous motifs and expression similarity is improved slightly (Pearson correlation: $R = 0.18$, $P$-value $\sim 0.013$; Spearman rank correlation: $\rho = 0.16$, $P$-value $\sim 0.025$).

Since many potential TF binding sites may not be functional, it would be beneficial to use transcription factor binding sites detected by large-scale chromatin immunoprecipitation (ChIP) experiments \[22\]. This complementary approach of determining functional interactions between TFs and upstream regions is not confounded by somewhat diverged motifs. We tested our result by using the dataset from Lee et al. (106 putative TFs from the ChIP experiment \[22\]). In this case we found a higher motif-expression correlation: Pearson correlation $R = 0.26$, $P$-value $\sim 0.002$; Spearman rank correlation: $\rho = 0.26$, $P$-value $\sim 0.001$.

It would be desirable to exploit the assumption that paralogs possessed the same motif structures and expression profiles soon after gene duplication. To avoid the substantial rate variation of amino acid distance among genes and the saturation of synonymous distance, we used the relative age of duplication; the unit is defined by the E.coli/yeast split, see BOX 1. We indeed have observed young duplicates have stronger correlation between expression similarity and motif similarity, while such correlation for ancient duplicate pairs disappears. For instance, using 0.9 unit as the cutoff, we divided the duplicate gene pairs into two same size groups and found that the coefficient of expression-motif correlation for the “recent” group is 0.19, while no correlation at all for the “ancient” group ($R \approx 0$). Note that this
pattern is insensitive to the cutoff value (not shown). Adjacent pairs of genes often show expression similarity even without sharing the same set of regulatory elements [23], because local chromatin structure may also modulate gene expression profiles. However, after excluding a few adjacent duplicate pairs in our analysis above, we obtain virtually the same pattern (not shown).

On the other hand, we have examined whether both the expression similarity and the fraction of shared motifs are negatively correlated with the age of gene duplication [16, 17, 24]. Fig. 1c shows a highly significant strong negative correlation between the expression correlation and the age of gene duplication (Pearson correlation: $R = -0.42, P < 0.0001$; Spearman rank correlation: $\rho = -0.44, P < 0.0001$). Similarly, Fig. 1d reveals a negative correlation between the fraction of shared motifs and the age of gene duplication (Pearson correlation: $R = -0.22, P = 0.001$; Spearman rank correlation: $\rho = -0.19, P = 0.008$). Using amino acid distance as an estimate of time of divergence gives virtually the same results (not shown).

One might be puzzled by the discrepancy between the strong co-expression pattern and relatively weak motif-expression correlation between duplicates. Actually, one can explain it in terms of how much of such co-expression can be explained by the correlation between expression and motif divergences. Recall that Yu et al. [15] claimed that 3.3% of gene pairs that share the same TF(s) are co-expressed, while our analysis showed that 7.0% of duplicate pairs are co-expressed. Meanwhile, the coefficient of correlation between expression divergence and regulatory motif divergence ($R = 0.15 - 0.18$) suggests that about $0.15^2 - 0.18^2 = 2.3\% - 3.2\%$ of expression variance can be explained by the regulatory motif divergence; even for the case of ChIP, it is only $0.26^2 = 6.7\%$. In this regard, the interpretation from these two studies is biologically consistent, while the significance level ($P$-value) may differ dramatically as the null hypotheses of two methods, as well as the sample size, are totally different. One possibility may include the non-linear co-expression property or the shared motif combinations rather than individual motifs [16, 25], that is, whether older duplicates have lower expression similarity because of having lost particular motif combinations. We have studied a more general linear or non-linear (e.g., logit) regression model, but no considerable improvement, partly due to the small sample size. One
solution is to study the motif-expression correlation under the phylogeny of a large gene family, which can be tested by using more sophisticated statistical methods [26].

Caveats notwithstanding, our analysis shows that experimental error is unlikely to be the only cause for the low explanatory power, and supports the notion that gene expression may be influenced by many trans-factors at different levels of gene networks. In addition to the cis-regulatory motif structure, other factors such as motif-motif interactions, transcription factor co-evolution, as well as the chromatin structure [25, 27, 28], may affect the expression levels without changing the regulatory motifs.

Roughly speaking, the weak motif-expression correlation between duplicates may be the consequence of two phenomena: (a) Duplicates with divergent motif structures actually have similar expression profiles, e.g. among the top 10 most highly co-expressed duplicate gene pairs, none of which has a fraction of shared motifs more than 50%. It could be due to the neutral turnover of binding sites, that is, randomly generated binding sites may cause the shift of transcription factor (TF) from its original one; or the convergent evolution, that is, the binding between different TF and their corresponding motifs may have similar impact on gene expression regulation. And (b) duplicates with similar set of regulatory motifs have low expression similarities, due to the context-dependence of trans-transcriptional regulation, which requires the combinatorial interaction of other bound proteins to function [25]. Further study is certainly needed when more data are available. On the other hand, evolution of regulatory motifs may not necessarily result in expression divergence because of the genetic robustness, or the alternative regulatory pathway in which the previous binding sites are not utilized. Besides, our analysis implies that ancient duplicate pairs may have more chances to be affected by the reshuffling of gene networks. Consequently, the expression-motif correlation may disappear because of the evolutionary reorganization of gene networks, resulting in considerable expression changes even the same motifs preserved [25]. Also worth mentioning is that there may be other trans-acting factors other than the transcription factors which have an influence on gene regulation [29].
If our interpretation is largely correct, some evolutionary models need to be revisited because the whole gene networks could be involved for shaping the expression divergence after gene duplication. For example, the duplication-degeneration-complementation (DDC) model of gene duplication [30] assumes that the status (presence or absence) of regulatory motifs (or motif modules) dominates the status of gene expression, which seems to be oversimplified according to our analysis; See [24] for other criticisms. The important information conveyed by our finding is that identification of motif structures is only one step in discovering the mechanism underlying gene expression regulation and evolution, and the effect of gene networks other than the cis-factors should not be neglected.
Acknowledgement: We appreciate Molley Vanhouten for helping collecting the microarray gene expression data. This work was supported by the NIH grant to X. G.
Box 1. Computational analysis for yeast genome datasets

Datasets:

(1) The complete sequences for 43 genomes of bacteria, archaea and \textit{S. cerevisiae} are available at NCBI website (http://www.ncbi.nlm.nih.gov/COG/), where gene families are classified as clusters of orthologous groups (COG). Amino acid alignments of 202 COGs containing 2 yeast duplicate genes were downloaded.

(2) The whole genome sequence of \textit{S. cerevisiae} was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/Ftp/index.html). The upstream regions (500bp) of each yeast gene in the COG family were extracted. According to the list of 50 "known" regulatory motifs compiled by Kellis \textit{et al.} [14] verified by experimentation (among those 55 motifs, 50 are unique motif sequences), we used Perl to scan along the upstream region of each yeast gene to identify all the motifs on both strands (BOX1-figure A). Besides, Kellis \textit{et al.} further predicted 72 "discovered" motifs, including 30 known motifs. The expression-motif (predicted) correlation is lower, due to the apparent prediction errors (not shown).

A. The distribution of identified motifs (number of motifs per gene) in 404 yeast genes in our study.
(3) A total of 276 microarray data points were collected from the on-line database [18]. As commonly suggested, we used the fold-change after the normalization for representing the gene expression level. Since we are focused on the expression divergence between duplication pairs, potential cross hybridization [31] may give us more conserved results.

(4) ChIP data [22] was downloaded from the Young lab’s homepage at http://web.wi.mit.edu/young/regulator_network. Each yeast gene in the 202 families was used as target gene to retrieve its corresponding TF(s) from the dataset.

B. The distribution of identified functional motifs in 404 yeast genes. About half of the yeast genes have no motif detected.

**Association study for co-expression pattern of duplicates**

The statistical significance of co-expression (top 1% of largest correlation coefficients of all possible gene pairs in the genome) between duplicate genes is calculated using a cumulative binomial distribution: 

\[ P_{c|c_0} = \sum_{c=c_0}^{N} \binom{N}{c} p^c (1-p)^{N-c}, \]

where \( N \) is the total number of gene pairs of interest, \( c_0 \) is the number of co-expressed gene pairs, and \( p \) is the probability of finding a co-expressed gene pair within the whole genome (see Yu et al. [15] for details).
Weighted correlation: The model for motif (M)-expression (E) regression can be written as

\[ E = M \beta + \alpha, \quad M = \sum_{k=1}^{K} \frac{\min(c_{1k}, c_{2k})}{\max(c_{1k}, c_{2k})}, \]

where \( c_{1k} (c_{2k}) \) is the number of copies of motif \( k \) in gene 1 (gene 2) of the duplicate pair. In particular, it becomes an unweighted correlation if \( c_{ik} = 1 \) if motif \( k \) is found in the upstream region of gene \( i \), or \( c_{ik} = 0 \) otherwise.

Identification of potential paralogous motifs

For each motif \( i \), the position in the upstream sequence of gene \( X \) is denoted by \( L_{ix} \). Two motifs located in the upstream regions of duplicates \( X \) and \( Y \) are paralogous if (1) they are the same type of motif; (2) the positions are sufficiently close, that is, \( |L_{ix} - L_{iy}| < W \), where \( W \) is the window size; and (3) when BLAST search with other yeast species [14], both upstream sequences around the binding sites show a high level of conservation respectively (not shown). Typically, we set \( W=50 \) bp. Several values of window sizes were considered. Although different window sizes gave different numbers of paralogous motifs, little difference was found when computing the correlation between the fraction of paralogous motifs and gene expression similarity (not shown).

Two-member gene families:

We conducted the following analysis for each gene family. First, we computed the fraction of shared motifs between duplicates. Second, we calculated the standardized expression covariance between duplicate genes. Third, we estimated the amino acid distance between duplicate genes using the Poisson correction. Since many gene duplications are very ancient, synonymous nucleotide distance may not be appropriate.

And fourth, to estimate the age of a gene duplication event [32], we reconstructed the phylogeny of the gene family using the neighbor-joining (NJ) method (MEGA2, http://www.megasoftware.net/). After carefully excluding the lateral gene transfer events, the linearized NJ tree allows us to compute the (average) duplication time relative to E.coli/yeast split, or the relative age of a duplication event (BOX1-figure B). Note that the yeast gene families used in our study include a significant portion of very ancient gene duplication events (figure B-1), while those events from the ‘ancient’ genome duplication
appear relatively recent in our analysis (figure B-2). Since the age estimation is crude, subject to the violation of molecular clock, lateral gene transfer, etc., we also used the amino acid distance as an alternative measure for the duplication time. Both measures gave consistent results and supported our main conclusion.

B. (1)
B. (1) Phylogenetic tree of the Purine nucleoside phosphorylase gene family (COG0005), showing an ancient yeast gene duplication. (2) Phylogenetic tree of the Galactokinase gene family (COG0153), showing a rather recent yeast gene duplication event. ♦ refers to the gene duplication between the yeast genes; ○ refers to the E.coli/yeast split.
References


18. On-line microarray datasets:
   http://www.genome.stanford.edu/yeast_stress
   http://cellcycle-www.stanford.edu
   http://genome-www.stanford.edu/zinc

CHAPTER V: GENERAL CONCLUSIONS

My Ph.D. research is mainly about applying statistical methods to the analyses of gene expression data, i.e. microarray data, putting the gene expression process into an evolution framework, and characterizing the expression evolution procedure. Such expression divergence analysis can deepen our understanding of the phenotypic evolution at the transcriptional level.

General Discussion

Molecular phylogeny currently plays major role in analyzing genomic data, trying to understand the relationship between genes, chromosomes and species. However, for another major source of genomic information, large-scale gene expression analysis, little research has been done from an evolutionary point of view. In chapter I, we reviewed a preliminary phylogenetic expression analysis developed by Gu (2000) that used a Brownian motion process to represent expression variation among duplicate genes in a gene family. The general Brownian-based model can be transformed and restricted to obtain several derived models or sub-models. Each sub-model can be applied to deal with specific biological questions, depending on the imposed restrictions.

Based on the E₀ model described by Gu (2004), we develop a fast algorithm to predict expression profiles at the ancestral nodes (genes). By comparing ancestral expression profiles with progeny expression profiles, so called expression divergence, expression profile changes along the duplication lineage, can be revealed and quantified. Such expression divergence can be used as an indicator of function divergence, showing if the gene activity is under selection pressure along that specific lineage and inferring the potential function difference between progeny genes. Details about ancestral expression inference can be found in chapter II.

The phylogenetic expression analysis proposed by (Gu 2004)¹ is rather complicated, especially as it requires the use of the maximum likelihood estimation which is sensitive to model assumptions. In chapter III, we transformed the key idea of the Brownian-based E₀

¹ Also see chapter I.
model into the form of an expression distance structure, and used the modified molecular phylogenetic approach to reconstruct an expression tree. Such expression phylogeny has the same convenience and flexibility as the molecular phylogeny in molecular evolutionary study. However, we did notice that there are some difference between expression phylogeny and molecular phylogeny, and such differences reveal the decoupling between expression profile evolution and sequence evolution.

In chapter IV, we used yeast expression data and motif data to study the relationship between expression divergence and motif divergence. Although it has long been believed that motif structure is the key factor in shaping the expression profiles, our analysis only reveals a weak coupling relationship between the two profiles. Many studies have shown that transcription regulation is a very complicated involving dynamic processes. It involves interactions between DNA and proteins, such as motifs and transcription factors; the interaction between proteins, such as the transcription factors and their cofactors; the modification and degradation of the proteins; the structure of the DNA sequence, such the condensation of local chromatin; etc. Our results simply indicate that motif structure is only part of the story and people should be very cautious when making assumptions about the relationship between motif structures and expression profiles.

**Future Research**

The research summarized in this dissertation is still in its theoretical stage. My next and immediate task is to apply those ideas to the analysis of real data. In order to accomplish this goal, statistical modeling and analysis in chapter II and III will be first incorporated into a program package which allows convenient and fast analysis, especially when dealing with massive datasets at the genomic level. In this research report, we showed some examples. In the future, we will extend these analyses to the whole genome of some organism, in particular, all the yeast gene families, and conduct the expression divergence analysis at the genome level.

During the study of motif and expression relationships, we realized the importance of gene networks in all aspects of the organism activities. In my next research project, I plan to combine the information of all the available components of gene networks, such as gene
expression, gene duplication, metabolic pathway, motif structure, null mutation mutants, etc., and see if I can identify some relationships among those components.

My long-term research goal is to integrate genomic data resources with evolutionary concepts and further investigate relationships among sequence divergence, expression divergence and function divergence. A key interest is to understand how those divergence processes can be related to or shaped by the structure and development of the gene networks.
ACKNOWLEDGEMENTS

I want to take this opportunity to thank all people who have helped me during my Ph.D. study. It is not possible for me to list all the names here and I apologize in advance if I omit someone’s name.

Special acknowledgment is due to my major advisor Dr. Xun Gu, who provided instruction, supervision and encouragement throughout my Ph.D. preparation. My whole thesis has benefited enormously from Dr. Gu’s insight in scientific research, and from his extensive knowledge in both biology and statistics. Next, I want to thank Dr. Kenneth J. Koehler for his insightful suggestion in statistical aspects. I would also like to thank other committee members Dr. Alicia L. Carriquiry, Dr. Drena Dobbs, Dr. Karin Dorman and Dr. Dan Nettleton for their advice and support.

My colleagues in the Computational Genomics and Evolution Lab accompanied me throughout my Ph.D. study and shared their research experience with me. I want to thank them for all their support and help. I also want to express appreciation to all the faculty members, students and staffs in both the Statistics Department and the Bioinformatics & Computational Biology program for providing me with such a professional research environment.

Finally, I want to particular thank my family back in China, especially my parents. Although they can not be around in my everyday life, it is their support and love that make all what I have accomplished so far possible.