Functional divergence and age distribution of vertebrate gene families

Yufeng Wang
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Genetics Commons

Recommended Citation
Wang, Yufeng, "Functional divergence and age distribution of vertebrate gene families " (2001). Retrospective Theses and Dissertations. 1089.
https://lib.dr.iastate.edu/rtd/1089

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.
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

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 bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI 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.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®
Functional divergence and age distribution of vertebrate gene families

by

Yufeng Wang

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

Major: Bioinformatics and Computational Biology
Major Professors: Xun Gu and Daniel Ashlock

Iowa State University
Ames, Iowa
2001

Copyright © Yufeng Wang, 2001. All rights reserved.
Graduate College
Iowa State University

This is to certify that the Doctoral dissertation of

Yufeng Wang

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 Major Program

Signature was redacted for privacy.

For the Graduate College
TABLE OF CONTENTS

GENERAL INTRODUCTION 1
   Introduction 1
   Dissertation Organization 4
   References 5

CHAPTER I: FUNCTIONAL DIVERGENCE IN CASPASE GENE FAMILY AND ALTERED FUNCTIONAL CONSTRAINTS:
STATISTICAL ANALYSIS AND PREDICTION 9
   Abstract 9
   Introduction 10
   Methods 11
   Results 17
   Discussion 24
   Acknowledgements 28
   Literature Cited 28

CHAPTER II: PHABA: A STATISTICAL APPROACH TO PREDICT FUNCTIONAL DIVERGENCE IN GENE FAMILIES FROM ALTERED FUNCTIONAL CONSTRAINTS 42
   Abstract 42
   Introduction 43
   Data and Methods 44
   Results 49
   Discussion 56
   Acknowledgements 59
   References 60

CHAPTER III: EVOLUTIONARY PATTERNS OF GENE FAMILIES GENERATED IN THE EARLY STAGE OF VERTEBRATES 74
   Abstract 74
   Introduction 75
GENERAL INTRODUCTION

Introduction

Biology is undergoing a revolution based on the accelerating determination of DNA sequences, including the complete genomes of a growing number of organisms (Adams et al 2000; International Human Genomes Sequencing Consortium, 2001; Venter et al., 2001). During this post-genomic era, functional genomics seeks to devise and apply technologies that take advantage of the flooding sequence information to analyze and predict in vivo functions of proteins (Doolittle 1996; McKusick 1997; Durbin et al. 1998). One of the missions of protein genomics is to make direct predictions on function(s) prior to biological experimentation.

Functional Divergence after Gene (Domain) Duplication

Gene duplication has long been thought to be a primary source of material for the origin of functional novelties (Ohno 1970; Sidow 1996). Similarly, domain shuffling (or domain duplication) is proposed to be a major evolutionary force to generate multi-domain proteins (Henikoff et al. 1997). In theory, gene duplication or/and domain shuffling events resulted in gene families and supergene families, in which several paralogous genes with sequence similarity employ related but distinct functions (Lundin 1993; Hughes 1994; Spring 1996). Under a conventional model, after gene (domain) duplication, one copy retains the original function and evolves under the functional
constraint, whereas the other copy is under relaxed selection constraint and may acquire novel function(s) by chance or may be silenced to a functionless pseudogene by deleterious mutations (Li 1983). The fate of duplicated genes has been predicted by a variety of specific models (Hughes 1994; Force et al. 1999; Lynch and Conery 2000). However, it is still unclear to what extent duplicated genes evolve under altered evolutionary constraints.

It is important to explore the functional divergence of a gene (supergene) family after gene duplication (domain shuffling). Current computational methods may provide a useful framework for further assays such as site-directed mutagenesis and protein engineering. For example, mechanisms of gene family tissue-specificity, functional liability vs. stability, enzyme-coenzyme specificity, ligand specificity have been addressed by using a similar phylogenetic analysis-mutagenesis-3D structure analysis approach (e.g. Trabesinger-Reuf et al. 1996; Yokoyama 1997). However, sequence similarity score, the fundamental criteria in these computational tools, may not be sufficient for detecting functional divergence (Eisen 1998). The difficulty is that most amino acid changes are neutral or nearly neutral, which are not directly related to functional innovations. Thus, the stochastic property in sequence evolution demands a probabilistic model (Gu 1999).

Critical Amino Acid Residues Responsible for Functional Divergence

For functional genomics, predicting critical amino acid residues responsible for functional divergence between homologous genes of a gene (supergene) family has great
potential. Comparing with time-consuming and labor-intensive mutagenesis screening, it can be very cost effective to first define targets in silico. Evidence from many case studies has clearly shown that only a few residues are truly involved in potential functional divergence after gene (domain) duplication (See Golding and Dean for a review). Indeed, given sequence data, it is challenging on how to identify these important residues from the background of evolutionary noise (Bork and Koonin 1998). Several algorithms have been proposed for predicting amino acid residues for experiments (Fetrow and Skolnick 1998; Montelione and Anderson 1999; Xia et al 2000; Irving et al, 2001). However, since they are usually based on some empirical rules rather than a statistical model, the reliability of predictions needs to be justified or at least just made explicit.

**Structural Basis of Functional Divergence after Gene (Domain) Duplication**

In the post-genomic era, one can expect more and more 3D structures of proteins will be determined due to high throughput structure determination approaches. In many cases, two homologous proteins with related but distinct functions have very similar 3D structures. On the other hand, several studies have shown that some residue changes that have dramatic consequences in physiology are not structurally important sites. It is interesting to investigate the interrelationship between functional divergence, protein structure and molecular evolution. Indeed, many amino acid changes caused by stochastic process are not related to functional divergence and thus do not result in structure changes.
Comparative Genomics of Gene (Supergene) Families

Traditionally, gene families in the human genome are thought to have arisen from several events: first, in the early stage of an animal lineage, multiple domain proteins or supergene families were generated by domain shuffling (duplication); second, two (?) rounds of genome duplications occurred in the early vertebrate stage, as a result, different tissue-specific isoforms were generated; third, local chromosome duplication may occur at a constant rate during the animal evolution; and fourth, some gene families have expanded recently. Comparative analysis is one of the most powerful methods available for understanding the diverse and complex biosystems, but it is often limited by a lack of comprehensive taxonomic sampling. Based on sequence homology, phylogenetic analysis provides a useful tool to build up a genome-level hierarchy of gene families.

One of the objectives of this study is to explore the evolutionary patterns in more than 2000 vertebrate gene families and to shed a light on the mechanism underlying genome complexity, for example, the origin of tissue specificity.

Dissertation Organization

The objective of this study is to develop and apply statistical methods to predict functional content from primary sequence and to explore the pattern of gene family evolution. This dissertation is composed of a general introduction, four chapters, each of which is in the journal manuscript format, and a general conclusions section. The four chapters that detail the core of the research work are outlined below.
Chapter 1 introduces a new statistical model for testing functional divergence and predicting critical residues (Gu 1999) by a case study in caspase gene family. By taking advantage of substantial experimental data of caspases, the functional/structural basis of our predictions are extensively studied. The objective of this study is to show the potential of combining new methodology with classical phylogenetic approach in functional genomics.

Chapter 2 extends the study to a comprehensive survey in functional divergence among a large number of gene families by using Gu (1999) method (PHYBA, phylogeny-based analysis). The technical issues, biological implications and potential applications are detailed addressed in this chapter.

Chapter 3 investigates the evolutionary patterns of 49 gene families that are generated in the early stage of vertebrates. The times of gene duplications are estimated to test the hypothesis of two-rounds (2R) of genome duplication. Complicated evolutionary patterns (2R/3R) are surveyed.

Chapter 4 examines the impacts of gene duplications on the functional divergence in vertebrate gene families. Two patterns of functional divergence after gene duplication(s) are illustrated.

References


Bork P, Koonin EV (1998) Predicting functions from protein sequences – where are the


Xia, Y., Huang, E.S., Levitt, M., & Samudrala, R. (2000). Ab initio construction of

CHAPTER I: 
FUNCTIONAL DIVERGENCE IN CASPASE GENE FAMILY AND ALTERED FUNCTIONAL CONSTRAINTS: STATISTICAL ANALYSIS AND PREDICTION

A paper accepted by Genetics

Yufeng Wang¹, Xun Gu¹,²

Abstract

In this paper, we explore the pattern of Type I functional divergence (i.e., altered functional constraints or site-specific rate difference) in the caspase gene family that is important for apoptosis (programmed cell death) and cytokine maturation. By taking advantage of substantial experimental data of caspases, the functional/structural basis of our posterior predictions from sequence analysis has been extensively studied. Our results are as follows. (1) Phylogenetic analysis shows that the evolution of major caspase-mediated pathways has been facilitated by gene duplications. (2) Type I functional divergence (altered functional constraints) is statistically significant between two major subfamilies, CED-3 and ICE. (3) 4 of 21 predicted amino acid residues (for site-specific

¹ Graduate student and Assistant professor, respectively, Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University.

² Author for correspondence
rate difference between CED-3 and ICE) have been verified by experimental evidence. And (4) we found that some CED-3 caspases may inherit more ancestral functions, whereas other members may employ some recently-derived functions. Our approach can be cost-effective in functional genomics to make statistically sound predictions from amino acid sequences.

**Introduction**

Gene family proliferation provides the raw material for functional innovation in higher eukaryotes. After gene duplication, the classical model (OHNO 1970) suggests that one gene copy maintains the original function, while the other copy is free to accumulate amino acid changes toward functional divergence. Since then, many more specific models have been proposed (e.g., LI 1983; CLARK 1994; FORCE et al. 1999). However, the details of functional divergence between duplicate genes remain largely unexplored. GU (1999) developed a method to detect amino acid residues that contribute to functional divergence after gene duplication, which can be considered as candidates for further experimentation. Its effectiveness for functional genomics needs to be tested by gene families with substantial biological/structural information such as caspase gene family.

Apoptosis, or programmed cell death, is an ordered process in which cells commit suicide when they are not needed or are potentially harmful. The key component in the apoptotic machinery is a cascade of caspases (cysteine aspartyl proteases). All caspases, initially inactive proenzymes, share the same processing scheme to achieve mature forms
after the cleavage(s) at specific Asp sites (KUMAR 1995; THORBERRY and LAZBNIK 1998). To date, at least 14 members of caspase gene family have been identified in mammals, which can be further classified into two major subfamilies, CED-3 (including caspases -2, -3, -6, -7, -8, -9, -10, and -14) and ICE (including caspases -1, -4, -5, -11, -12, and -13) (NICHOLSON and THORNBERY 1997). Substantial evidence has shown that the CED-3 type caspases are essential for most apoptotic pathways (YUAN et al. 1993; KUIDA et al. 1996). In contrast, the major function of the ICE type caspases is to mediate immune response, although some members may play a role in cell death in some circumstances (YUAN and HORVITZ 1990; WANG et al. 1998). X-ray crystallography has also shown a significant structural difference between these two types of caspases (e.g., WILSON et al. 1994; ROTONDA et al. 1996).

In this paper, we take advantage of experimental evidence of caspases to study the functional-structural basis of statistical predictions from GU’s method (1999). We first statistically evaluate the functional divergence between CED-3 and ICE subfamilies. We then show our predictions for critical amino acid residues are consistent with the observations from protein 3-D structure or functional assay. Our analysis shows the potential of evolutionary approach for functional genomics.

Methods

The data set

We conducted an exhaustive search (e.g., the Gapped BLAST and PSI-BLAST) in several major databases to find all available sequences that are homologous to C.
elegans CED-3 gene. After synthetic peptides, ESTs, partial sequences and redundant sequences were removed, the final data set includes 42 CED-3 homologous protein sequences, whose accession numbers are: (1) Casp-3: U13737 (human 3-a), U13738 (human 3-b), U49930 (Rat 3-a), U58656 (Rat3-b), Y13086 (mouse), U27463 (hamster), AF083029 (chicken), D89784 (frog); (2) Casp-7: U37448 (human), Y13088 (Mouse), AF072124 (Rat), U47332 (hamster); (3) Casp-6: U20536 (human), AF025670 (Rat), Y13087 (mouse), AF082329 (chicken); (4) Casp-8: AF102146(human), AF067841(mouse); (5) Casp-10: U60519 (human 10a), U86214 (human 10/b), AF111345 (human 10/d); (6) Casp-9: U60521 (human); (7) Casp-2: U13021 (human), U77933 (Rat), Y13085 (mouse), U64963 (chicken); (8) Casp-14: AF097874 (human), AJ007750 (mouse); (9) Casp-1: X65019 (human), AF090119(horse), L28095 (mouse), U14647 (Rat), D89783 (frog ICE-A), D89785 (frog ICE-B); (10) Casp-4: Z48810 S78281 (human); (11) Casp-5: X94993 (human); (12) Casp-13: AF078533 (human); (13) Casp-11: Y13089 (mouse); (14) Casp-12: Y13090 (mouse); (15) Invertebrate caspase: P42573 (C. elegans CED-3), Y12261 (Drosophila melanogaster), U81510 (armyworm, Spodoptera frugiperda).

Multiple Alignment and Phylogenetic Analysis

The multiple alignment of 42 caspase amino acid sequences was obtained by the program CLUSTALX (THOMPSON et al. 1997), followed by manual editing according to the structure information (NICHOLSON and THORNBERY 1997). A phylogenetic tree was inferred by the neighbor-joining method (SAITO and NEI 1987) using
MEGA2.0 ([http://www.megasoftware.net/](http://www.megasoftware.net/)). PAUP4.0 and PHYLIP were used to examine whether the inferred phylogeny is sensitive to any tree-making method. To evaluate the intensity of functional constraints in each caspase, we calculated the ratio of nonsynonymous to synonymous rates between human/mouse orthologs using LI (1993) and modified NEI and GOJOBORI (in MEGA2.0) methods.

**Type I functional divergence (altered functional constraint) analysis**

*Types of amino acid configurations*

Consider a multiple alignment of a gene family with two homologous genes A and B (Figure 1A). Although different classifications have been put forward (e.g., LIVINGSTONE and BARTON 1996), we adopt the following schemes: *(i)* *Type 0* represents amino acid configurations that are universally conserved through the whole gene family, implying that these residues may be important for the common function shared by all member genes. *(ii)* *Type I* represents amino acid configurations that are very conserved in gene A but highly variable in gene B, or *vice versa*, implying that these residues may have experienced altered functional constraints. *(iii)* *Type II* represents amino acid configurations that are very conserved in both genes but their biochemical properties are very different, e.g., charge positive vs. negative, implying that these residues may be responsible for functional specification in the different subfamilies. And *(iv)*, amino acid configurations at many residues are not such clear-cut that they have to be regarded as unclassified *(Type U)*.
Several algorithms have been proposed to define these types of amino acid configurations automatically (e.g., CASARI et al. 1995; LICHTARGE et al. 1996; LANDGRAF et al. 1999). However, these methods are subject to various problems, e.g., negligence of phylogenetic tree, unclear statistical basis, or arbitrary cut-off for classification. Thus, a statistical model seems useful to deal with these problems.

**Functional divergence and altered functional constraint**

After gene duplication, two duplicates can undergo substantial functional divergence. It seems that only a small portion of residues that are involved in functional divergence (GOLDING AND DEAN 1998). The trajectories of differentiation can affect the evolutionary pattern of the gene family divergence in several ways. According to GU (1999), Type I functional divergence refers to the evolutionary process that results in altered functional constraints (or evolutionary rate) between two duplicate genes, regardless of the underlying evolutionary mechanisms. Intuitively, type I amino acid configuration is likely to be observed at the residue with different evolutionary rates between duplicate genes (Type I functional divergence). However, because of the stochastic nature of molecular evolution, each site, no matter whether it is related to functional divergence, has a non-zero probability of becoming any type of amino acid configuration (Figure 1B). Therefore, instead of classifying ad hoc Type I amino acid configuration, GU’s (1999) method is to compute the (posterior) probability of Type I functional divergence for each amino acid site. Note that we can define Type II functional divergence in the same manner.
Statistical modeling for Type I functional divergence

It is conceptually convenient to use the ancestral gene (before duplication) as a reference. For each duplicate gene cluster, the evolutionary rate at a site may differ from the ancestral gene, which is called F₁-site (functional divergence-related); otherwise it is called F₀-site (functional divergence-unrelated). As shown in figure 1B, different evolutionary rates between duplicate genes are expected only when a site being F₁ at least in one cluster (e.g., sites 3, 4, and 5 in panel B); a status denoted by S₁. The coefficient of type I functional divergence (θ) between two gene clusters is defined as the probability of a site being status S₁, i.e., θ = P(S₁). The alternative status is S₀, which means a site being F₀ in both clusters (i.e., the evolutionary rate of each duplicate gene is the same as the ancestral gene, e.g., sites 1 and 2 in Figure 1B). Obviously, P(S₀) = 1 - θ. The null hypothesis is θ = 0, which means that the evolutionary rate is virtually the same between duplicate genes (as well as the ancestral gene) at each site.

Let λₐ and λₐ be the evolutionary rates of a site in clusters A and B, respectively, which vary among sites. For a site being F₀ in both clusters (status S₀) with a probability of 1 - θ, we can assume λₐ = λₐ without loss of generality. However, for a site being S₁ (i.e., being F₁ at least in one cluster) with a probability of θ, we have λₐ ≠ λₐ. To avoid too many parameters, GU (1999) made the following simplification: Under S₁, although λₐ > λₐ at some sites, and vice versa at others, over all sites λₐ and λₐ are statistically independent. Remember that λₐ = λₐ under S₀ simply means a complete correlation between λₐ and λₐ. Figure 2 outlines the statistical procedure on how to estimate θ from sequences.
Prediction of critical amino acid residues

If $\theta > 0$ significantly, it provides statistical evidence that Type I functional divergence (altered functional constraint) may have occurred after gene duplication. It is of interest to predict which residues are responsible for the altered functional constraints after gene duplication. This can be achieved by the posterior analysis (Figure 2). Let $P(S_i|X)$ be the posterior probability of a site being $S_i$ when the amino acid configuration $(X)$ is observed. Since the alternative status $S_0$, with posterior probability $P(S_0|X)=1-P(S_1|X)$, means no altered functional constraint, the predicted residues are only meaningful when $P(S_i|X) > 0.5$ such that the posterior odd ratio $R(S_i/S_0) = P(S_i|X)/P(S_0|X) > 1$. A more stringent cut-off may be $P(S_i|X) > 0.67$, or $R(S_i/S_0) > 2$.

Cluster-specific Type I functional divergence: functional distance analysis

The two-cluster analysis described above cannot tell in which gene cluster the altered functional constraint took place after gene duplication. This problem can be solved by a simple method when at least three homologous gene clusters are available. For any cluster $i$, let $\theta_i = P_i(F_i)$ be the probability of a site having a different rate from the ancestral gene, and $P_i(F_0) = 1-\theta_i$ be the probability of having the same rate. Consider two clusters $i$ and $j$ that the coefficient of Type I functional divergence is denoted by $\theta_{ij} = P_{ij}(S_j) = 1- P_{ij}(S_0)$. If a site being $F_1$ or $F_0$ is independent between clusters, we have the relation $P_{ij}(S_0) = P_i(F_0) \times P_j(F_0)$, or, $1-\theta_{ij} = (1-\theta_i)(1-\theta_j)$. Therefore, we define Type I functional distance between clusters $i$ and $j$ as $d_F(i,j) = -\ln (1-\theta_{ij})$, and functional branch
length for cluster $i$ or $j$ as $b_F(i) = -\ln (1-\bar{\theta}_i)$ and $b_F(j) = -\ln (1-\bar{\theta}_j)$, respectively.

Obviously, $d_F(i,j)$ is additive, i.e.

$$d_F(i,j) = b_F(i) + b_F(j).$$

When the coefficient of Type I functional divergence ($\bar{\theta}_{ij}$) for each pair of clusters is estimated, the matrix of $d_F(i,j)$ can be easily computed. Then, a standard least square method is implemented based on Eq. (1) for estimating all $b_F$’s. A large $b_F$ indicates substantial altered functional constraints in this gene cluster, while $b_F = 0$ indicates that the evolutionary rate of each site in this duplicate gene is almost identical to the ancestral gene. In other words, a duplicate gene cluster with $b_F = 0$ may contain a larger component of ancestral function, compared to other gene clusters.

Results

Evolution of caspase mediated molecular pathways

The phylogenetic tree

The evolutionary tree (Figure 3) of caspase gene family was inferred by the neighbor-joining (NJ) method with Poisson distance (SAITOU and NEI 1987). Parsimony method (PAUP4.0) and likelihood method (PHYLIP) give virtually the same topology (data not shown). The presence of caspases in vertebrates, arthropods and nematodes suggests that the emergence of caspase gene family might be close to or even earlier than the origin of animal kingdom.

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

*Evolutionary innovations of caspase-mediated apoptosis pathway by gene duplications*

To understand the origin of different caspase-mediated biochemical pathways in apoptosis, we compared the evolutionary relationship of (CED-3 type) caspases with apoptotic pathways (Figure 4). Our major finding is that major evolutionary lineages of caspases may coincide with different caspase-mediated apoptotic pathways triggered by specific death signals. That is, (i) Caspase-9 is a key component in the mitochondrial initiated pathway which is initiated by the intracellular stimuli, upstream Bcl-2, and Apaf-1 proteins (BUDIHARDJO et al. 1999). (ii) Caspase-2 initiates the apoptosis induced by negative signaling after B cell Ag receptor (BCR) ligation (CHEN et al. 1999). (iii) Apoptosis mediated by caspases -8 and -10 are similar, both initiated by responding to the death receptors (DRs), which contain the death effector domain (DED).
(iv) Uniquely, caspase-14 is not processed by any known death stimuli (VAN DE CRAEN et al. 1998). In summary, since ancient origins, these caspases may evolve through different avenues and provide cells a potential to initiate apoptosis in response to a variety of intracellular or intercellular stimuli.

Interestingly, although upstream initiator caspases (I-Casps, e.g., Casp -2, -9, -8/-10) are recruited by different receptors under different physiological or pathological stimuli, they all eventually catalyze the same set of downstream effector-caspases (caspases -3, -6, -7), which are the real killers that commit the cell suicide (Figure 4). Our results suggest (1) that gene duplication followed by functional divergence is one major mechanism to generate the complexity of apoptotic network; and that (2) such a process is constrained by coordinated regulation. Indeed, in the last step, effector-caspases as real killers remain unchanged when more initial death signals are continuously recruited at different levels during the evolution of apoptotic pathways.

Predicting critical residues for type I functional divergence (altered functional constraints) between CED-3 and ICE subfamilies

We have shown that the altered functional constraint between these two subfamilies is statistically significant: $\theta = 0.29 \pm 0.05$ [the ML option in GU's method (1999)]. Further, we use the posterior probability $P(S_i|X)$ to predict critical amino acid residues responsible for Type I functional divergence between CED-3 and ICE subfamilies (Figure 2). The baseline of the site-specific profile measured by $P(S_i|X)$ is around 0.2-0.3 (Figure 5A). Thirty-two sites (16% of total sites) have $P(S_i|X) > 0.5$. The
fact that most sites have less than 50% scores indicates their similar functional roles between CED-3 and ICE.

Although the posterior analysis is widely used in bioinformatics, the cut-off value is always subject to somewhat arbitrary. We found that when the first 21 highest-scored residues are removed from the multiple alignment, the estimate of $\theta$ is virtually 0. These 21 amino acid residues (among 198 residues) corresponding to the cut-off value $P(S|X) > 0.61$ are then chosen for further analysis. This empirical procedure is meaningful only when $\theta > 0$ significantly.

*The functional-structural basis of altered functional constraints*

We have mapped these 21 predicted sites onto the 3D structure of caspases. The resolved X-ray crystal structures of human caspases -1 and -3 (WILSON et al. 1994; ROTONDA et al. 1996) have been used to illustrate the structural features of ICE and CED-3 subfamilies, respectively. From the literature, we have found experimental evidence for four predicted residues that are involved in the functional-structural divergence between CED-3 and ICE subfamilies (Figure 5B):

(i) Residue 161: critical for CED-3 caspase substrate specificity by interacting with a unique surface loop in 3D structure (ROTONDA et al. 1996).

Amino acid residue 161 (hCasp-1 348)\(^3\) is predicted with $P(S|X)=0.998$. At this position, all 22 sequences from the CED-3 subfamily contain an invariant tryptophan (W), whereas a variety of residues are present in the ICE subfamily (Figure 6). Rotonda et al. (1996) have shown an extra surface loop that is only present in CED-3 type

\(^3\)#1(hCasp-1 #2):#1 is the alignment position, #2 corresponds to the primary sequence of human casp-1.
caspases (see the blocked region in Figure 6). Crystal structural analysis reveals that W348 is a key determinant for the caspase-3 (CED-3) specificity. First, W348 forms a narrow pocket with the surface loop that is highly conserved in the CED-3 subfamily (Figure 7). The steric constriction due to this pocket determines the preference of caspase-3 to the substrates with small hydrophilic side chains. Second, W348 along with a group of residues forms a hydrogen bond network that affects the interaction with the substrate. Therefore, any change in W161 (hCasp-1 348) of the CED-3 type could cause breakdown of the hydrogen bond network with the substrate, altered 3D structure, and altered enzymatic activity. Consequently, all CED-3 type caspases maintain this invariant W161 (hCasp-1 348), due to a very strict functional constraint.

In contrast, the surface loop shared with CED-3 caspases seems to be deleted in all ICE-type caspases as shown in the block in Figure 6. Instead of a pocket-like geometry in CED-3 type, a shallow depression is formed in human caspase-1 (ICE) at the homologous position (Figure 7). The orientation and configuration of residues in this vicinity can be well changed by the substrate preference of the CED-3 type to hydrophilic side chains, while the ICE type to hydrophobic ones. Hence, the relaxed evolutionary constraint observed at this position in the ICE subfamily is likely to be caused by the 3D structural difference.

(ii) Residues 86 \([P(S_i|X) = 0.75]\) and 88 \([P(S_i|X) = 0.74]\): responsible for 3D difference with unknown functional role

The altered functional constraints at positions 86 \([P(S_i|X) = 0.75]\) and 88 \([P(S_i|X) = 0.74]\) are consistent with the 3-D structural difference between two
subfamilies. In human caspase-1, Residues 86 (hCasp-1 249) and 88 (hCasp-1 251) appear to lie in an extra loop that is not present in the CED-3 type caspases (WILSON et al. 1994; ROTONDA et al. 1996). Although the functional role of this extra small loop is unclear, one may expect that this intrinsic radical structural difference may lead to functional divergence between CED-3 and ICE subfamilies.

(iii) Residue 131 \( P(S|X) = 0.866 \): proteolytic site specific to the ICE subfamily

All caspases are synthesized as inactive proenzymes that need to be processed to the mature forms (NICHOLSON et al. 1995). However, distinct cleavage sites within the precursors are found for two subfamilies. D131 is known as a cleavage site in human caspase-1 (ICE type) (THORNBERRY et al. 1992). All ICE type caspases preserve an Asp (D) at this position, except for mouse caspase-12 (Asn, E). However, human caspase-3 (CED-3 type) utilizes two other Asn sites for cleavage (ROTONDA et al. 1996) so that the functional role of position 131 in CED-3 caspases is no longer important. Therefore, the altered evolutionary constraints at this position can be well explained by the different utilization of cleavage sites for the precursor processing between CED-3 and ICE subfamilies.

**Pattern of Type I Functional divergence among CED-3 type caspases**

The CED-3 subfamily consists of a specific group of caspases that mediate the programmed cell death in a well-regulated proteolytic cascade and employ related but distinct functions. Here we address an interesting problem, i.e., to infer the trend of altered functional constraint of each cluster.
We study five gene clusters: caspases -3, -7, -6, -8/-10, and -2. Due to insufficient data, caspase-9 was excluded, and caspases -8 and -10 are grouped for their closely related function (FERNANDES-ALNEMRI et al. 1996). The upper diagonal of Table 1 shows pairwise coefficients of Type I functional divergence (θ) between them; all of them are significantly larger than 0 (p < 0.05), with only one exception, i.e., θ = 0.006 between caspase-7 and cluster -8/-10.

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

Long functional branch lengths (b_F) of caspases -3, -6 and -2 suggest that these genes may have undergone extensive altered functional constraints, as a result of specialized functional roles in apoptosis (Figure 8). Supportive experimental evidence can be summarized as follows: (i) The non-redundant functional role of caspases -3 in neurological apoptosis has been confirmed by Caspase-3 -/- knockout mice (KUIDA et al. 1996). (ii) Caspase-6 and -3 have different substrate specificity, but both participate in protease amplification cycle by activating each other, which triggers a series of apoptotic interactions (LAZEBNIK et al. 1995, SRINTVASULA et al. 1996). (iii) Caspase-2 has its unique dual-role position in positive and negative regulation in apoptosis, by differential expression of two alternative splicing isoforms (2_L and -2_S) (WANG et al. 1994). This
dual-role property has been confirmed by knockout mice: caspase-2 deficiency cause one defective apoptotic pathway (mediated by granzyme B and perforin), but accelerate another one (cell death of motor neurons) (BERGERON et al. 1998).

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

For each duplicate gene, the average intensity of functional constraints can be approximately measured by the $d_N/d_S$ ratio between appropriate orthologous sequences (e.g., human-mouse). Interestingly, caspases -3, -6 and -2 (long $b_F$) have lower $d_N/d_S$ ratios than caspases -7 and -8/10 (zero $b_F$), indicating that Type I functional divergence in caspases may result in a stronger functional constraint (Figure 8B).

Discussion

In this paper, we have investigated the pattern of Type I functional divergence (altered functional constraint) during the evolution of caspase gene family. Our results are as follows. (1) Phylogenetic analysis shows that the evolution of major caspase-mediated pathways has been facilitated by gene duplications. (2) Type I functional divergence is statistically significant between CED-3 and ICE subfamilies. (3) Among 21 critical amino acid residues predicted to be responsible for Type I functional divergence between CED-3 and ICE, four of them have been verified by experimental evidence. And (4) based on the functional distance analysis, a hypothesis is proposed that caspases (-7,
and -8/-10) may largely maintain the ancestral function of CED-3 subfamily, whereas caspases -3, -6, and -2 have highly diverged.

As shown before, Type I functional divergence refers to the evolutionary process that results in altered functional constraints (or evolutionary rate) between two duplicate genes, regardless of the underlying evolutionary mechanisms. The significance of our study is two-fold: First, we have shown that altered functional constraint after gene duplication may play an important role in evolutionary novelties after gene duplication. Second, the site-specific profile based on posterior analysis is useful for designing a cost-effective approach in functional genomics, e.g., the strategy for a large-scale mutagenesis.

Predicted sites for Type I functional divergence without evidence could be either lacking of experimental data, or due to statistical artifacts (e.g., cutoff value). On the other hand, experimentally verified sites but missed by our analysis could be other types of functional divergence (e.g., Type II). In practice, the cut-off value should be weighted by other biological information. And the validity of the posterior-based analysis holds only when \( \theta > 0 \) significantly.

It is clear that the accuracy of our prediction depend on how strong is the association between functional divergence and altered functional constraints. This proposition can be traced back to FITCH’s (1970) covarion theory but has not been well understood. Another factor affecting the prediction accuracy is the complexity of the model. Although GU’s (1999) method is simple and fast, it can be improved by using a
more complicated model but the computational time may become a real problem (GU 2001).

Our model does not explicitly consider the underlying mechanism. Thus, the method of SUZUKI and GOJOBORI (1999) dealing with positive selection on single sites could perform better in such a case when natural selection is the major force for functional diversity and the divergence time is short so that the synonymous rate can be well calculated. However, there is a significant portion of gene families in vertebrate genomes that have been generated in the early stage of vertebrates (~500 million years ago) (WANG and GU 2000). For these important datasets, application of SUZUKI and GOJOBORI’s (1999) methods seems unfeasible because synonymous substitutions are almost certainly saturated. Nevertheless, there are many other methods available for the functional prediction from molecular evolutionary analysis (e.g., see GOLDING AND DEAN 1998 as a review), and the new-developed methodology is complement to most of them. An appropriate combination can greatly advance our research.

In the gene duplication theory, ancestral function of a gene family is conceptual rather than measurable. Functional distance analysis provides a quantitative measure for the altered functional constraints between the ancestral gene and one duplicate gene. This evolutionary measure has raised an interesting hypothesis that caspases -7 and -8/-10 may represent the function of the common ancestor of the CED-3 subfamily since their respective functional branch lengths are virtually zero. We hope this hypothesis can be tested by experiments with a combination of ancestral sequence inference (GOLDING and DEAN 1998).
As one of important mechanisms for functional or structural divergence, co-evolution between residues can occur in the early stage after gene duplication (POLLOCK et al. 1999; NAYLOR and GERSTEIN 2000). By contrast, in the late stage (within each gene cluster), protein evolution is likely to be under simple purifying selection so the occurrence of co-evolution between residues may not be very frequent. Since GU's (1999) method is to model altered functional constraints in the late stage, the assumption of site-independence seems not very unrealistic.

Similar to any site-specific analysis, our prediction is sensitive to the quality of the multiple alignment. We have examined the multiple alignment of caspase family, particularly in the surrounding regions of these four verified predicted sites (Figure 6). To the best of our knowledge, the alignment can be considered "nearly-optimized". For example, the alignment of position 161 is almost indisputable.

In conclusion, we have conducted a case-study to show the capability of predicting Type I functional divergence (i.e., altered functional constraints) from sequence evolution. Moreover, our analysis has shown that a comprehensive approach including various computational methods and multi-level information (from sequence to experimental data) could be tremendously beneficial for understanding functional diversity of a large gene family in the post-genomics era.
Acknowledgements

The study is support by the NIH grant RO1 GM62118 to X. G. We are grateful to Drs. C-I Wu and Gavin Naylor for constructive comments, which have improved the manuscript significantly. Thanks Jianying Gu for assistance.

Literature Cited


LAZEBNIK, Y. A., A. TAKAHASHI, R. D. MOIR, R. D. GOLDMAN, G. G.


POLLOCK, D., W. R. TAYLOR, and N. GOLDMAN, 1999 Coevolving protein


SRINITVASULA, S. M., T. FERNANDES-ALNEMRI, J. ZANGARILLI, N.


VAN DE CRAEN, M., G. VAN LOO, S. PYPE, W. VAN CRIEKINGE, I. VAN DEN


Table 1: θ values and d_F values from pair-wise comparisons in the CED-3 subfamily.

<table>
<thead>
<tr>
<th></th>
<th>Caspase 3</th>
<th>Caspase 7</th>
<th>Caspase 6</th>
<th>Caspase 8/10</th>
<th>Caspase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>θ_{ij} ± Se</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>d_F(i,j) ± Se</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Caspase 3</td>
<td>0.437 ± 0.178</td>
<td>0.844 ± 0.113</td>
<td>0.467 ± 0.113</td>
<td>0.540 ± 0.110</td>
<td></td>
</tr>
<tr>
<td>Caspase 7</td>
<td>0.574 ± 0.257</td>
<td>0.579 ± 0.198</td>
<td>0.006 ± 0.022</td>
<td>0.198 ± 0.184</td>
<td></td>
</tr>
<tr>
<td>Caspase 6</td>
<td>1.858 ± 0.724</td>
<td>0.865 ± 0.470</td>
<td>0.527 ± 0.190</td>
<td>0.627 ± 0.125</td>
<td></td>
</tr>
<tr>
<td>Caspase 8/10</td>
<td>0.629 ± 0.212</td>
<td>0.006 ± 0.022</td>
<td>0.749 ± 0.401</td>
<td>0.306 ± 0.180</td>
<td></td>
</tr>
<tr>
<td>Caspase 2</td>
<td>0.777 ± 0.239</td>
<td>0.221 ± 0.229</td>
<td>0.986 ± 0.335</td>
<td>0.365 ± 0.259</td>
<td></td>
</tr>
</tbody>
</table>

Note:

θ_{ij} : the coefficient of the functional divergence between clusters i and j,

d_F(i,j): the distance of the functional divergence between clusters i and j,

d_F(i,j) = -\ln (1-θ_{ij}), where i and j are the row and column designation numbers respectively, and i≠j.

Se: standard error.
Figure 1. (A) Classification of amino acid configurations for two duplicate gene clusters. Type 0 sites represent these universally conserved through the whole gene family. Type I sites are these very conserved in one cluster but highly variable in the other. Type II sites are these very conserved in both clusters but with very different biochemical properties. And Type U sites are these unable to be classified. (B) A diagram shows the stochastic nature of molecular evolution. Each site (represented as a box) has a non-zero probability for any type of amino acid configuration. At sites 1 or 2, no altered functional constraint occurs in both clusters, a status defined as $S_0 = (F_0, F_0)$. At sites 3, 4, or 5, altered functional constraint occurs at least one cluster, a status defined as $S_0 = (F_1, F_0)$ or $(F_0, F_1)$ or $(F_1, F_1)$ (see Method for the details).
Input: aligned amino acid sequences of two clusters (A, B) and the phylogeny

Site-specific profile

Posterior analysis:
P(S|X) = \theta f(X|S) / f(X)

Probabilistic model of a site in each cluster:
f(X_A|\lambda_A), f(X_B|\lambda_B), where X_A and X_B are amino acid configurations in A and B. In the fast algorithm of Gu (1999), X_A and X_B are simplified to the expected number of substitutions (Gu and Zhang 1997) so that f(X_A|\lambda_A) and f(X_B|\lambda_B) are Poisson processes. See Gu (2001) for a formal likelihood treatment.

Conditional joint probability: assume rates \lambda_A and \lambda_B varies among sites according to a gamma distribution.
Under S_0, \lambda_A = \lambda_B = \lambda, so
f(X|S_0) = E[f(X_A|\lambda) f(X_B|\lambda)]
Under S_1, \lambda_A, \lambda_B independent, so
f(X|S_1) = E[f(X_A|\lambda_A)]E[f(X_B|\lambda_B)]
where X = (X_A, X_B) and E for expectation.

Figure 2. A flow chart to illustrate GU's (1999) method.
Figure 3. The phylogenetic tree of the caspase gene family, inferred by the neighbor-joining method based on the amino acid sequence with Poisson correction. Bootstrap values >50% are presented. Initiator caspases (I-Casps) are involved in upstream regulatory events, and effector caspases (E-Casps) directly lead to cell disassembly.
Figure 4. A schematic of evolution of caspase-mediated pathways. Note that the ancestral function of caspases (as well as the origin ICE type caspases) is uncertain. A, B, C correspond to ancestral nodes in Figure 3. Bcl-2/Apaf, BCR, death receptors (DRs), TNFR1, CD95 are death signals for specific apoptotic pathways. Caspases -3/-6/-7 are effector caspases (E-Casps) which are the real killer proteins in the programmed cell death.
Table 5. (A) The site-specific profile for predicting critical amino acid residues responsible for the functional divergence between CED-3 and the ICE subfamilies, measured by the posterior probability of being functional divergence-related at each site \( P(S_i | X) \). The arrows point to four amino acid residues at which functional divergence between two subfamilies has been verified by experimentation. (B) Four predict sites that have been verified by experimentation.
Figure 6. Alignment of predicted regions of caspases. Four predicted sites with experimental evidence are highlighted. The sites with asterisks are predicted residues within this region. The blocked region in the c-terminus is the critical region for CED-3 substrate specificity: most CED-3 type caspases form a surface loop, whereas a shallow depression is found in ICE-type caspases.
Figure 7. upper: core structure component of human caspase-3 chain A and chain B (pdb:1pau)(ROTONDA et al. 1996); lower: core structure components of human caspase-1 chain A and chain B (pdb: 1ICE)(WILSON et al. 1994). Among 16 predicted sites, 13 of them are mapped to chain A and chain B, three sites are in the cleaved fragment not shown in the structures. The structural images were created by the RASMOL program (SAYLE and MILNER-WHITE 1995).
Figure 8. (A) A star-like topology of the CED-3 caspases in terms of Type I functional branch length $D$. Biological evidence of functional specification for each caspase cluster is shown in the blocked boxes. (B) Functional branch length ($b_F$) and the ratio of nonsynonymous to synonymous rates ($d_N/d_S$) for each gene cluster, which were computed by using human-mouse sequences.
CHAPTER II:
PHYBA: A STATISTICAL APPROACH TO PREDICT
FUNCTIONAL DIVERGENCE IN GENE FAMILIES FROM
ALTERED FUNCTIONAL CONSTRAINTS

A paper to be submitted to the Journal of Molecular Biology

Yufeng Wang¹,², Jianying Gu¹ and Xun Gu¹,³

Abstract

It is important to understand the functional divergence among homologous clusters in a gene family. Predicting such functional divergence from sequence information is one of the challenging tasks in functional genomics. The difficulty stems from the evolutionary "noise", which needs a stochastic model to deal with. In this paper, we present a statistical method of modeling and testing Type I functional divergence (altered functional constraints), and further predicting critical amino acid

¹ Graduate students and Assistant Professor, respectively, Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University.

² Primary researcher and author.

³ Author for correspondence
residues. The potential and limitation of our methodology (PHYBA, phylogeney-based analysis) in functional genomics has been discussed by examples.

**Introduction**

With the recent completion of the human genome sequences, the door has open wide on the era of functional genomics (International Human Genomes Sequencing Consortium, 2001; Venter et al., 2001). Understanding gene family evolution and finding functional diversity among member genes from primary sequences is one of major issues in functional genomics, because the relationship between gene family diversity and the origin of multicellularity, tissue-specificity or body plan has been experimentally demonstrated. It is believed that gene duplication plays important role in the generation of gene families: after gene duplication, one duplicate gene receives most selection pressure, whereas the other copy diverges and is free to accumulate amino acid changes to achieve novel function, the functional constraints at some sites are expected to change (Ohno, 1970; Li, 1983). Despite that the general concept of gene duplication followed by functional divergence has been widely accepted and supporting data are available from functional assays, it is still unclear how to measure the degree of functional divergence between two duplicate genes (Golding & Dean, 1998). The difficulty stems from the fact that most amino acid changes observed are not directly related to functional divergence. For instance, amino acid differences between two duplicate genes can be the result of an ancient gene duplication or a more recent rapid divergence. Current function prediction approaches based on homologous search,
phylogenetic analysis, structural inference may not be sufficient to distinguish between two possibilities (e.g., Shortle et al., 1998; Ortiz et al., 1999; Xia et al., 2000; Koonin et al., 2001; Irving et al., 2001).

Here, we present a statistical approach to detect functional divergence between homologous gene clusters. The probabilistic model filters the amino acid changes related to functional divergence from the background "noise" which mainly represents neutral evolution (Gu, 1999; Gu, 2001). The principle of our method is based on the premise that functional divergence after gene duplication is highly correlated with the change of evolutionary rates, i.e., altered evolutionary constraints can be an indicator for functional divergence. We shall describe the framework of this approach and its potential applications in functional genomics.

**Data and Methods**

**Statistical Testing for Type I functional divergence (altered functional constraints)**

Consider two homologous gene clusters that are given rise by a gene duplication. In the early stage of the gene duplication, amino acid substitutions responsible for functional diversification between duplicate genes can be driven by functional redundancy, positive selection, coevolution of residues, and other forces. As a result, in the late stage, the functional constraints at some residues differ between two clusters, and the evolutionary altered between two clusters. This type of functional divergence is defined as Type I, characterized as altered functional constraints as well as altered
evolutionary constraints between homologous gene clusters (Figure 1A) (Gu 1999). For other types of functional divergence, one can refer to Wang and Gu (2001). The statistical modeling and testing for Type I functional divergence is composed of three major parts (Figure 1B).

1. Estimating the coefficient of Type I functional divergence ($\theta$)

Given a multiple alignment, the amino acid configuration of a site is denoted by $X = (X_1, X_2)$. The transition probability matrix for a particular period of time $t$ is given by $P = e^{\lambda R T}$, where the rate matrix $R$ is empirically determined by the Dayhoff model (Dayhoff et al., 1978). The evolutionary rate ($\lambda$) varies among sites according to a gamma distribution, where the shape parameter $\alpha$ describes the strength of rate variation among sites. The conditional probability of observing $X$ at a site, $f(X|\lambda)$, can be obtained from the Markovian property, given a phylogenetic tree.

Under a simple two-state model, each site in one cluster can have two possible states, $F_0$ (functional-divergence-unrelated) and $F_1$ (functional-divergence-related). There are four possible combinations for two clusters: $(F_0, F_0)$, $(F_0, F_1)$, $(F_1, F_0)$, and $(F_1, F_1)$. The state $(F_0, F_0)$ means no altered functional constraint at a site in both clusters, resulting in the same evolutionary rate, i.e., $\lambda_1 = \lambda_2$. In contrast, in the rest of three states, the evolutionary rates of two clusters are independent, because such sits has experienced altered functional constraints at least in one of the clusters. We define two non-degenerate states as $S_0$ and $S_1$. $S_0 = (F_0, F_0)$, and $S_1 = (F_0, F_1) \cup (F_1, F_0) \cup (F_1, F_1)$. The joint probability of amino acid configuration $X = (X_1, X_2)$ conditional on State $S_0$ or $S_1$ can be written as
\[ F(X|S_0) = \int_0^\infty f(X_1|\lambda) f(X_2|\lambda) \phi(\lambda) d\lambda = \mathbb{E}[f(X_1|\lambda) f(X_2|\lambda)], \]

\[ F(X|S_1) = p(X_1) p(X_2) = \mathbb{E}[f(X_1|\lambda_1)] \times \mathbb{E}[f(X_2|\lambda_2)], \]

respectively.

We define the coefficient of functional divergence as \( \theta = P(S_1) \). Thus the joint probability of observing amino acid configuration \( X \) is given by

\[ p(X) = (1-\theta)f(X|S_0) + \theta f(X|S_1). \]

Under the assumption of site-independence, the likelihood function is given by

\[ L = \prod_{\text{site}} p(X). \]

\( \theta \) can be estimated by a fast algorithm (Gu 1999) or a standard likelihood approach (Gu 2001).

2. Predicting critical amino acid residues for Type I functional divergence: A Bayesian approach

Rejection of the null hypothesis (i.e., \( \theta > 0 \) significantly) provides statistical evidence for altered functional constraints after gene duplication. Further, we can use a site-specific profile \( (Q_k) \) to predict critical amino acid residues, which is the posterior probability of \( S_1 \) at site \( k \),

\[ Q_k = \theta \frac{f(X|S_1)}{p(X)}, 0 \leq Q_k \leq 1. \]

Large \( Q_k \) indicates a high possibility that the functional constraint of a site is different between two clusters.

3. Functional distance analysis: mapping functional divergence to each individual gene cluster

For a gene family with multiple clusters, we are interested in detecting in which duplicate gene the altered functional constraint took place and. We proposed a simple method to solve this problem (Wang and Gu 2001). By definition, \( \theta_i = P_i(F_1) \) is the probability of a site being \( F_1 \) in cluster \( i \); \( P_i(F_0) = 1 - \theta_i \) is the probability of a site being \( F_0 \).
For any two clusters $i$ and $j$, the pair-wise coefficient of Type I functional divergence is defined as $\theta_{ij} = P_{ij}(S_1) = 1 - P_{ij}(S_0)$. Given independence of states $F_0$ and $F_1$, we have $P_{ij}(S_0) = P_i(F_0) \times P_j(F_0)$, i.e., $1 - \theta_{ij} = (1 - \theta_i)(1 - \theta_j)$. Thus, we can define *Type I functional distance* between clusters $i$ and $j$ as $d_F[i,j] = -\ln (1 - \theta_{ij})$, and *functional branch length* for cluster $i$ or $j$ as $b_F[i] = -\ln (1 - \theta_i)$ and $b_F[j] = -\ln (1 - \theta_j)$. Obviously, $d_F[i,j]$ is additive, i.e., $d_F[i,j] = b_F[i] + b_F[j]$

The matrix of $d_F[i,j]$ can be easily obtained after the ML estimation of $\theta_{ij}$. The least square method can be implemented to estimate the Type I functional distance ($b_F$). A large $b_F$ indicates this duplicate gene has undergone substantial altered functional constraints. On the other hand, an essential zero $b_F$ value indicates almost identical evolutionary rate of each site in this duplicate gene to the ancestral gene, suggesting this duplicate copy may contain a large component of ancestral function.

4. *Software (PHYBA)*

To explore residues responsible for functional divergence we have developed a new tool entitled PHYBA (PHYlogeny-Based Analysis). Developmental emphasis of PHYBA has been centered on four main concerns: correctness of results, ease of use, expandability, and accessibility. PHYBA has been developed to function identically whether the environment is Microsoft Windows 98/NT or Unix. Operation of PHYBA requires the user to supply an alignment, a gene tree (a NJ tree can be generated by PHYBA), and an amino acid substitution matrix (standard ones are provided). Also, if available, a suitable protein structure can be provided by a file in PDB format. After PHYBA has preformed the statistical analysis of the sequences, the user may examine
important residues for functional divergence by the posterior analysis, which can be plotted onto the alignment or the protein structure (if available). By providing the ability to interact with the protein structure, new discoveries about the interrelationships, such as spatial clustering of residues implicated in functional divergence may be found.

**Data sets**

*Collection of 29 gene families with two homologous clusters*

29 two-cluster gene families were obtained from GenBank, EMBL, and Hovergen (http://pbil.uni-lyon1.fr/). Human homologues were obtained by searching against the Celera human genome sequence database (http://public.celera.com/index.cfm). Invertebrate ourgroups were obtained by the Gapped BLAST and PSI-BLAST searches (Altschul et al. 1997). Partial sequences and redundant sequences were removed from further analysis.

*Collection of domain sequences in Jak gene family*

We searched the pfam database (http://pfam.wustl.edu/) for sequences that have tandem kinase (JH1) or pseudokinase (JH2) domains of Jak genes. The kinase domain sequences of FGFR and EGFR gene families were also downloaded.

*Collection of protein sequences of MyoD gene family*

The amino acid sequences of vertebrate MyoD gene family (MyoD, Myf-5, MRF-4, MyoG) were obtained from the databank hovergen. The Gapped BLAST and PSI-BLAST searches were performed to find the invertebrate homologues by using human MyoD gene as a query sequence (Altschul et al., 1997).

The accession numbers of sequences used in this study are available upon request.
**Multiple Alignment and Phylogenetic Analysis**

The multiple alignment of 29 gene families and MyoD gene family were obtained by the program Clustal X (Thompson et al. 1997). The multiple alignment of tandem kinase and pseudokinase domains of Jak was obtained from pfam, followed by manual editing. Phylogenetic tree was inferred by the neighbor-joining method (Saitou and Nei 1987) using the software MEGA2.0 (http://www.megasoftware.net/). Parsimony and likelihood methods were also used for phylogenetic reconstruction to examine if the tree topology is robust to the tree-making method.

**Results**

**Altered evolutionary/functional Constraints: a common pattern in protein evolution**

We have conducted a large-scale statistical analysis on the changes of evolutionary/functionary constraints in 29 vertebrate two-cluster gene families generated by gene duplication (Table 1).

First, we tested whether the evolutionary rates are constant among amino acid sites in each gene cluster, assuming the rate variation follows a gamma distribution, where the shape parameter \( \alpha \) specifies the range of variation (Gu and Zhang, 1997). The respective estimates of \( \alpha \) \( (\alpha_1 \text{ and } \alpha_2) \) for gene cluster 1 and cluster 2 in each family are presented in Table 1. Most of gene clusters have low \( \alpha \) values, suggesting different sites have different evolutionary/functional constraints.
The next interesting question is after gene duplication, whether two duplicated copies maintain the same functional constraint or at least one of them have some altered constraint. We used the coefficient of functional divergence ($\theta$) as an index for measuring functional divergence (Gu 1999). Table 1 shows that $\theta$ values between two gene clusters vary from 0.032 to 0.669, all of which are significantly greater than 0 except for TAP and Adenosine receptor 2 gene family whose $\theta$ value are essentially zero. This implies that altered functional constraint is one of general patterns of functional divergence after gene duplication. Indeed, a recent analysis including more datasets has confirmed this conclusion, i.e., most duplicates have undergone altered functional constraints after gene duplications (data not shown).

**Prediction of critical amino acids responsible for functional divergence in homologous genes -- Jak gene family as an example:**

The second component of our method is to predict critical amino acid residues contributing to the functional divergence from sequence information. Recently, we have shown the potential of posterior prediction in several case studies (e.g., Wang and Gu, 2001). Here we report the preliminary study of functional divergence between two homologous domains in Jak gene family.

Jak (Janus kinase) is a non-receptor tyrosine kinase, which plays important roles in signal transduction pathways (Darnell et al. 1994; Leonard and O’Shea 1997). In mammals, Jak gene family is composed of four member genes: Jak1, Jak2, Jak3 and Tyk2. A unique feature of the Jak family is the existence of an additional domain called pseudokinase(JH2) domain. As the kinase domain (JH1) has all of the features of a
typical tyrosine kinase domain, the function of the pseudokinase domain is unclear, while pseudokinase domain may have lost the catalytic activity, as lacking of some key residues (Velazquez et al. 1995; Luo et al. 1997). Some experimental evidence has shown abnormal phenotypes when some residues are mutated in the pseudokinase domain or the whole region is deleted in the knock-out mice (Guschin et al. 1995; Feng et al. 1997; Zhou et al. 1997).

To explore the evolutionary pattern of JH1 and JH2 domains, we reconstructed a NJ tree, including Jaks and two close related protein tyrosine kinases (PTK), FGFR and EGFR. The multiple alignments of kinase domains are available upon request. Phylogenetic analysis shows that tandem kinase (JH1) and pseudokinase (JH2) domains are evolutionarily distinct (Figure 2). Indeed, tandem kinase domains (JH1) in Jaks appear close to the functional kinase domains of FGFRs and EGFRs, while the pseudokinase domains (JH2) of Jaks form a unique clade. Actually we found that the JH2 domain had been generated before the emergence of major PTK supergene families (data not shown).

It is generally believed that the pseudokinase domains (JH2) have lost the catalytic activity but may acquire some new functions. To test whether functional constraints have been shifted at least at some residues during JH1/JH2 domain evolution, we estimated the coefficient of Type I functional divergence (θ) between them. The θ values is estimated to be $\theta = 0.412 \pm 0.049$ between tandem kinase (JH1) and pseudokinase (JH2) domains in Jak proteins, providing statistical evidence for supporting the hypothesis of altered functional constraints between JH1 and JH2.
A site-specific profile based on the posterior probability $Q_k$ is used to identify critical amino acid residues that are responsible for the altered functional constraints between tandem kinase (JH1) and pseudokinase (JH2) domains. Among 212 amino acid residues, 154 (73%) residues have $Q_k < 0.5$ which apparently have little contribution. For the remaining 58 amino acid residues with $Q_k > 0.5$, 21 of them show very high probability of being functional divergence related [$Q_k > 0.9$]. Without uncertainty, these 21 residues can be grouped into two categories: (1) conserved in the tandem kinase (JH1) domain, but variable in the pseudokinase (JH2) domain; (2) conserved in the pseudokinase domain, but variable in the tandem kinase domain (Table 2).

**Category I:** For 12 residues belonging to this category I, site 137 [$Q_{137} = 0.957$] has been demonstrated a determining site for the function of tandem kinase (JH1) domain, corresponding to the second tyrosine (highlighted Y) of a conserved (E/D)YY motif in Jak2 protein. This motif, which is located in the activation loop of Jak2, regulates the kinase activity by autophosphorylation (Feng et al. 1997). In Tyk2, these two consecutive tyrosines (YY) have also been identified as phosphorylation sites (Gauzzi et al. 1996).

Interestingly, the multiple alignments clearly show that residue 137 is invariant in the tandem kinase (JH1) domain. In contrast, the same position in pseudokinase domains (JH2) shows a variety of amino acids with very different chemical properties. (Figure 3A). This observation can be easily explained as a relaxed selective constraint which was caused by loss-of-function in phosphorylation in JH2 domains.
**Category II:** Nine predicted residues belong to category II (Figure 3B). Among them, site 103 is predicted to be highly functional-divergence related \[ Q_{103} = 0.954 \]. Experimental data shows that a glutamic acid (E)-to-lysine (K) substitution occurring at this site in the pseudokinase (JH2) domain hyperactivated Drosophila and mammalian Jak-Stat pathway (Luo et al. 1997). It seems likely that after the domain duplication, the tandem kinase domain (JH1) largely maintains the original catalytic function, while the pseudokinase domain (JH2) may achieve some unidentified new functions, resulting in a set of JH2-specific conserved residues.

**Association of the functional distance with the phenotypic pattern of transgenic mice — MyoD gene family as an example:**

Functional distance analysis, an important component of our statistical approach can provide a statistical measure for the extent of functional divergence contributed by individual clusters in a multi-gene family. Here we report the inference from the sequence evolution of MyoD gene family, which is strikingly consistent with the knock-out mice assays.

MyoD gene family, an essential component in myogenesis, is comprised by four member genes: MyoD, Myf5, MRF-4, and MyoG (Gerhart and Kirschner, 1997). This gene family exhibits an interesting feature: its four members form a hierarchical gene network for muscle formation, where MyoD and Myf-5 play at the top of the hierarchy, and MRF-4 and MyoG play downstream (Figure 5A). Thus, they are called myogenic master regulatory factors. It has been widely accepted that MyoD gene family arose by two successive rounds of gene duplication in the early vertebrate stage, though the exact
duplication time is under debate (Atchley et al., 1994, Hughes, 1999). The neighbor-
joining tree shows that the first round of gene duplication took place prior to the
divergence of Urochordata and Vertebrata, which gave rise of the ancestors of MyoD/
Myf-5, and that of MRF-4/MyoG, and the second round of gene duplication occurred in
the early vertebrate stage, prior to the emergence of bony fishes (Figure 4), which
generated four member genes. It was proposed that novel functions have been developed
after these two gene duplications (Sidow, 1996; Force, 1998).

Based on the phylogeny, we computed the pair-wise coefficients of functional
divergence (θ) for MyoD gene family (the upper triangle of Table 3). The maximum
likelihood (ML) estimates of all θ values are significantly greater than 0 (p < 0.05),
suggests that certain level of functional diversity did occur among four gene clusters. For
example, θ is estimated to be 0.213 ± 0.054 between MyoD and Myf-5, 0.171 ± 0.110
between MRF-4 and MyoG, implying altered functional constraints among products of
the second round of gene duplication.

To further decompose the respective contribution of each gene cluster to the
overall functional divergence, we performed the functional distance analysis (see
Methods). The pairwise functional distances (d_F) between clusters are shown in the lower
triangle of Table 3. The star-like tree presented in Figure 5B shows the Type I functional
branch length of each cluster (b_F), estimated by the least-squared method. The null
hypothesis of equal b_F for each cluster has been statistically rejected (p<0.05). Because
functional branch length of a duplicate gene cluster is a measure for the difference of
functional constraint from the ancestral gene (see Method), long functional branch
lengths \( (b_F) \) of MyoD and MRF-4 suggest that these genes may have undergone extensive altered functional constraints, as a result of specialized functional roles, while virtually zero functional branch lengths of Myf-5 and MyoG may inherit a large component of ancestral functions with unchanged functional constraints. These results are supported by the phenotypic patterns of knock-out mice (Table 3): (1) Myf-5 \( (b=0.020) \) is of less functional divergence than MyoD \( (a=0.22) \). This provides a reason why Myf-5, rather than MyoD, is capable of compensating the functional roles of both MyoD and MRF-4 (Mohun, 1992; Zhang et al. 1995). We speculate that Myf-5 may represent the original function of common ancestors in this myogenesis hierarchy. (2) MyoG \( (d=0.043) \) is less functionally divergent than MRF-4 \( (c=0.128) \). This is consistent with the transgenic results: MyoG employs an essential component from the ancestral function so that MyoG null mutants die, whereas MRF-4 may largely represent derived function so that knock-out mice exhibit normal phenotypes (Nabeshima et al., 1993; Hasty et al., 1993). (3) Moreover, the functional divergence between MyoD/Myf-5 and MRF-4/MyoG is statistically significant \( (e=0.204) \), reflecting the fact that the biofunctional role (receive inductive signals and trigger myocyte differentiation) of their common ancestor has already split after the first round of gene duplication. However, Myf-5 may still remain bi-functional.

With the functional distance and knock-out result, we propose a simple model illustrating the functional divergence and consequent phenotypic specification in each gene cluster (Fig 5C). The common ancestor carried on dual functions \((i^*m^*)\): respond to the inductive signals (denoted by \( i \)) and initiate myocyte differentiation (denoted by \( m \)).
After the first round of gene duplication, one duplicate (the ancestor of Myf-5/MyoD) preserved ancestral functions, while the other copy (the ancestor of MRF-4/MyoG) underwent the altered functional constraint and lost the sensibility to inductive signals (i^+m'). The diversification of MyoD member genes after the second round of gene duplication is more complicated: (1) Myf-5 largely inherited the bi-functional role (i^+m^+) whereas MyoD eventually only function in the upstream of hierarchy (i^+m^—). (2) On the other hand, MyoG preserved most of its ancestral function with some derived indispensable function of myocyte differentiation (i^m^*) based on the fact that the null homozygotes of MyoG die; MRF-4 acquired some new novel function and had escaped from the purifying selection (i^M^—). This proposed model supports the conventional conjecture that the two gene copies after gene duplications exhibit altered functional constraints, one of which seems to have inherited the ancestral function, and the other has developed new function (Li 1983).

Discussion

In this paper, we have shown the principle, design and applications of a statistical method for predicting functional divergence in homologous genes. Here we discuss a few technical issues, biological implications and potential applications.

Technical issues:

It should be noted that to assure accuracy and lower statistical errors for estimating θ, large sample size (i.e., number of sequences, N) is more applicable. It has been mathematically proven that the variance of estimated θ, Var(θ), is largely affected
by number of sequences tested (Gu 1999). Increasing the number of sequences results in fewer statistical errors. Although it is still computable with $N<4$, $N\geq 4$ is appropriate for each gene cluster (Table 1).

On the other hand, the quality of alignment, which is the prerequisite for good phylogenetic reconstruction, is important for the estimation because our method is essentially based on the reliable tree topology. It has been shown that unreliable alignment will result in substantially different estimate for $\theta$ (J. GU et al., unpublished data). We did not include those partial sequences shorter than 80% of full length in this analysis because: (1) after the complete gap deletion of alignment, the total number of sites drop dramatically, as a result, the inferred phylogenetic tree is usually less accurate; (2) some deleted sites due to the partial fragment(s) may actually contain informative residues for functional divergence, thus excluding those fragments is worthwhile.

It is important to test the prediction power of Type I functional divergence. To examine the accuracy of a method, one has to find a testing dataset with known functional domains/motifs or 3D structures, and then calculate the percentage of recovery. However, this testing strategy may not be applicable because there does not exist such perfect testing dataset. Nevertheless, it would be possible to develop some quantities to measure the relation between statistical prediction and known functional information.

Gu’s (1999) method applies only to the Type I functional divergence, which resulted in altered functional constraints among homologous genes within a gene family. For example, in Jak gene family, the amino acid residue 137, which has the highest
posterior probability of being functional-divergence related, shows a typical Type I pattern, i.e. invariant tyrosine in all tandem kinase domains, whereas highly variable in pseudokinase domains. On the other hand, Type II functional divergence results in no altered functional constraints but changes in amino acid properties. We can observe a cluster-specific pattern of amino acid types in Type II functional divergence: highly conserved in all gene clusters, while the conserved residues are different in different clusters. For instance, at position 13, all sequences preserve a glycine in the tandem kinase domain, whereas all sequences have a threonine in the pseudokinase domain. It is possible that some residues that were not detected by the present method may have undergone Type II functional divergence without any observable altered functional divergence. A statistical method for Type II functional divergence is under the way (Gu, unpublished results).

**Biological implications and potential applications**

The first attempt to connect functional divergence with functional constraints was made by Fitch and Markowitz (1970). It is postulated in their covarion theory that a nucleotide or amino acid site is either variable or invariable, and the site that is invariable in one cluster after speciation or other evolutionary events can be variable. This "covarian behavior" (Gaucher et. al., 2001) is indeed a special case of Type I functional divergence defined by Gu (1999). Our two-state \((F_0/F_1)\) model provides a statistical-computational framework for the (evolutionary) rate-(functional) pattern change. The preliminary results from the study of 29 gene families show that a shift of functional constraints after gene duplication is a general phenomenon in the evolution of novel
functions. Moreover, the coefficient of functional divergence (θ) is highly gene-specific, and not strongly correlated with the average sequence diversity or the divergence time (data not shown).

Our newly developed statistical method may have potential for biomedical and pharmaceutical research in the era of functional genomics. For example, one can use our posterior site-specific profile to identify amino acid residues that are responsible for functional divergence. This may provide a guideline for functional assay design. Moreover, mapping predicted residues onto 3-D protein structures will help enhance the understanding of function/structure relationship in gene families.

Another application of our statistical approach may be related to the phenotypic pattern prediction. Phenotypes in knock-out mice among homologous genes indicate the consequence of functional diversity in the context of gene architecture. The underlying biological mechanism is so complicated that it is difficult to predict the phenotypic pattern of knock mice via site-by-site experimentation. Nevertheless, the clear connection between the altered functional constraints and the phenotypic changes in gene function [e.g., MyoD, caspase (Wang and Gu, 2001), and Jak (J. Gu et al., 2001) gene families] provide some evolutionary perspective for detecting phenotypic patterns.

**Acknowledgement** This study was supported by NIH grant ROI GM62118 to X. G. We are grateful to Kent Vender Velden who developed the graphical interface for PHYBA.
References


proteins using covarion-based evolutionary approaches: Elongation factors. *Proc Natl
Acad Sci USA.* 98: 548-552.

Gauzzi, M.C., Celazquez, L., McKendry, R., Mogensen, K., Fellous, M. & Rellegrini, S.
(1996). Interferon-α-dependent activation of Tyk2 requires phosphorylation of positive

MA.


Gu, J., Wang, Y., & Gu, X. Evolutionary Perspectives for functional divergence of
Jak gene family. (submitted to J. Mol. Evol.)


Guschin, D., Rogers, N., Briscoe, J., Witthuhn, B., Watling, D., Horn, F., Pellegrini, S.,
tyrosine kinase Jak1 in the Jak/Stat signal transduction pathway in response to


Table 1: Coefficient of functional divergence (\(\theta\)) for 29 gene families with two homologous clusters.

<table>
<thead>
<tr>
<th>Gene Family</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>#sites</th>
<th>(\alpha_1)</th>
<th>(\alpha_2)</th>
<th>(\theta \pm \text{Se}^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAP</td>
<td>TAP1(20)</td>
<td>TAP2(21)</td>
<td>516</td>
<td>2.170</td>
<td>2.217</td>
<td>0.032±0.078</td>
</tr>
<tr>
<td>Adenosine receptor 2</td>
<td>ADR2A(5)</td>
<td>ADR2B(4)</td>
<td>318</td>
<td>0.302</td>
<td>0.584</td>
<td>0.134±0.142</td>
</tr>
<tr>
<td>Aminolevulinate synthase</td>
<td>ALAS1(5)</td>
<td>ALAS2(5)</td>
<td>442</td>
<td>0.246</td>
<td>0.358</td>
<td>0.163±0.073</td>
</tr>
<tr>
<td>Endothelin receptor</td>
<td>ENDRA(6)</td>
<td>ENDRB(6)</td>
<td>384</td>
<td>0.416</td>
<td>0.181</td>
<td>0.224±0.078</td>
</tr>
<tr>
<td>transferrin</td>
<td>LTF(8)</td>
<td>TF(5)</td>
<td>595</td>
<td>0.776</td>
<td>0.567</td>
<td>0.234±0.055</td>
</tr>
<tr>
<td>Multi-drug resistance protein</td>
<td>MDR1(8)</td>
<td>MDR(3)</td>
<td>1252</td>
<td>0.419</td>
<td>0.263</td>
<td>0.252±0.080</td>
</tr>
<tr>
<td>Bone morphogenetic protein</td>
<td>BMP2(8)</td>
<td>BMP4(7)</td>
<td>444</td>
<td>0.502</td>
<td>0.443</td>
<td>0.283±0.067</td>
</tr>
<tr>
<td>Y-box binding protein</td>
<td>YB1a(13)</td>
<td>YbvarA(5)</td>
<td>179</td>
<td>0.524</td>
<td>0.041</td>
<td>0.314±0.164</td>
</tr>
<tr>
<td>Cholecystokinin receptor</td>
<td>CCKRA(4)</td>
<td>CCKRB(9)</td>
<td>400</td>
<td>0.299</td>
<td>0.508</td>
<td>0.340±0.127</td>
</tr>
<tr>
<td>Clathrin light chain</td>
<td>CLTA(8)</td>
<td>CLTB(6)</td>
<td>206</td>
<td>0.135</td>
<td>0.095</td>
<td>0.362±0.235</td>
</tr>
<tr>
<td>Natriuretic peptide</td>
<td>ANP(10)</td>
<td>BNP(6)</td>
<td>85</td>
<td>1.013</td>
<td>1.864</td>
<td>0.406±0.124</td>
</tr>
<tr>
<td>Parvalbumin/oncomodulin</td>
<td>PVAL(7)</td>
<td>OMD(5)</td>
<td>108</td>
<td>1.139</td>
<td>0.838</td>
<td>0.436±0.182</td>
</tr>
<tr>
<td>Topoisomerase II</td>
<td>Topo2A(6)</td>
<td>Topo2B(4)</td>
<td>1497</td>
<td>0.292</td>
<td>0.105</td>
<td>0.439±0.055</td>
</tr>
<tr>
<td>cyclooxygenase</td>
<td>COX1(8)</td>
<td>COX2(11)</td>
<td>583</td>
<td>0.499</td>
<td>0.403</td>
<td>0.455±0.079</td>
</tr>
<tr>
<td>N/R-cadherin</td>
<td>N-CAD(7)</td>
<td>R-CAD(4)</td>
<td>737</td>
<td>0.388</td>
<td>0.381</td>
<td>0.457±0.062</td>
</tr>
<tr>
<td>ISDBA/NTCP</td>
<td>ISDBA(5)</td>
<td>NTCP(4)</td>
<td>336</td>
<td>0.505</td>
<td>1.107</td>
<td>0.462±0.163</td>
</tr>
<tr>
<td>P-type ATPase 7</td>
<td>ATP7A(4)</td>
<td>ATP7B(4)</td>
<td>1379</td>
<td>0.500</td>
<td>0.362</td>
<td>0.472±0.088</td>
</tr>
<tr>
<td>ELFT/FUT9</td>
<td>ELFT(4)</td>
<td>FUT9(4)</td>
<td>330</td>
<td>0.877</td>
<td>2.187</td>
<td>0.489±0.116</td>
</tr>
<tr>
<td>LAMP</td>
<td>LAMP1(6)</td>
<td>LAMP2(5)</td>
<td>375</td>
<td>1.360</td>
<td>1.206</td>
<td>0.492±0.065</td>
</tr>
<tr>
<td>SERCA</td>
<td>SERCA1(13)</td>
<td>SERCA2(5)</td>
<td>990</td>
<td>0.068</td>
<td>0.235</td>
<td>0.492±0.090</td>
</tr>
<tr>
<td>Activin receptor type II</td>
<td>ActR2A(10)</td>
<td>ActR2B(6)</td>
<td>500</td>
<td>0.411</td>
<td>0.441</td>
<td>0.501±0.086</td>
</tr>
<tr>
<td>FML receptor</td>
<td>FMLR(7)</td>
<td>C5AR(8)</td>
<td>283</td>
<td>0.623</td>
<td>0.802</td>
<td>0.522±0.134</td>
</tr>
<tr>
<td>Crystalline α</td>
<td>CRYA-A(16)</td>
<td>CRYA-B(9)</td>
<td>78</td>
<td>0.823</td>
<td>0.768</td>
<td>0.527±0.244</td>
</tr>
<tr>
<td>Corticotropin releasing factor</td>
<td>CRF(7)</td>
<td>Urotensin(5)</td>
<td>112</td>
<td>0.770</td>
<td>1.778</td>
<td>0.529±0.141</td>
</tr>
<tr>
<td>XDH/AOX</td>
<td>XDH(6)</td>
<td>AOX(4)</td>
<td>1319</td>
<td>0.480</td>
<td>0.539</td>
<td>0.550±0.053</td>
</tr>
<tr>
<td>Heme oxygenase</td>
<td>HO1(6)</td>
<td>HO2(8)</td>
<td>245</td>
<td>1.179</td>
<td>0.983</td>
<td>0.560±0.157</td>
</tr>
<tr>
<td>Endothelin</td>
<td>END1(6)</td>
<td>END2(6)</td>
<td>130</td>
<td>0.707</td>
<td>1.713</td>
<td>0.563±0.168</td>
</tr>
<tr>
<td>Serum albumin</td>
<td>ALB(13)</td>
<td>AFP(6)</td>
<td>507</td>
<td>0.984</td>
<td>2.318</td>
<td>0.565±0.075</td>
</tr>
<tr>
<td>Bradykinin receptor</td>
<td>BKR1(4)</td>
<td>BKR2(5)</td>
<td>313</td>
<td>1.120</td>
<td>0.871</td>
<td>0.669±0.120</td>
</tr>
</tbody>
</table>

Note:
1: \(\text{Se}\) represents standard error,
2: the number in parenthesis is the number of sequences in each gene cluster.
Table 2: $\theta_{ij}$ values from all the combinations of the pair-wise comparisons for Jaks, FGFRs, and EGFRs.

<table>
<thead>
<tr>
<th></th>
<th>JH1</th>
<th>JH2</th>
<th>FGFR</th>
<th>EGFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>JH1</td>
<td>0.412±0.049</td>
<td>0.222±0.055</td>
<td>0.496±0.071</td>
<td></td>
</tr>
<tr>
<td>JH2</td>
<td>0.531</td>
<td>0.566±0.080</td>
<td>0.561±0.080</td>
<td></td>
</tr>
<tr>
<td>FGFR</td>
<td>0.251</td>
<td>0.835</td>
<td>0.432±0.096</td>
<td></td>
</tr>
<tr>
<td>EGFR</td>
<td>0.685</td>
<td>0.823</td>
<td>0.566</td>
<td></td>
</tr>
</tbody>
</table>

Note:
JH1: the tandem kinase domain in Jak; JH2: the pseudokinase domain in Jak.
The upper diagonal shows the $\theta_{ij}$ of all the pair-wise comparisons for Jaks, FGFRs, and EGFRs. $\theta_{ij}$ is defined as the coefficient of the functional divergence between cluster i and cluster j. The lower diagonal shows the $d_F[i,j]$ value which is defined as the distance of the functional divergence between cluster i and cluster j, and $d_F[i,j] = -\ln(1-\theta_{ij})$, where i is the row designation number, and j is the column designation number, and $i \neq j$. 
Table 3: θ values and d_F values from pair-wise comparisons in the MyoD subfamily.

<table>
<thead>
<tr>
<th></th>
<th>MyoD</th>
<th>Myf-5</th>
<th>MRF-4</th>
<th>MyoG</th>
</tr>
</thead>
<tbody>
<tr>
<td>MyoD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myf-5</td>
<td>0.240 ± 0.120</td>
<td>0.316 ± 0.097</td>
<td>0.226 ± 0.083</td>
<td></td>
</tr>
<tr>
<td>MRF-4</td>
<td>0.541 ± 0.192</td>
<td>0.380 ± 0.195</td>
<td></td>
<td>0.171 ± 0.110</td>
</tr>
<tr>
<td>MyoG</td>
<td>0.496 ± 0.223</td>
<td>0.256 ± 0.142</td>
<td>0.188 ± 0.151</td>
<td></td>
</tr>
</tbody>
</table>

Note:
θ_{ij} : the coefficient of the functional divergence between clusters i and j,
d_F[i,j]: the distance of the functional divergence between clusters i and j,
d_F[i,j] = -ln (1-θ_{ij}), where i and j are the row and column designation numbers, respectively, and i≠j.
Se: standard error.
Table 4: Phenotypes of transgenic mice in which myogenic genes are knocked out.

<table>
<thead>
<tr>
<th>Gene(s) nullified</th>
<th>Phenotype</th>
<th>Compensation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MyoD</td>
<td>Normal</td>
<td>Myf-5 increases</td>
</tr>
<tr>
<td>Myf-5</td>
<td>Normal</td>
<td>Not applicable</td>
</tr>
<tr>
<td>MyoDMyf-5</td>
<td>Die</td>
<td>No</td>
</tr>
<tr>
<td>MyoG</td>
<td>Die</td>
<td>No</td>
</tr>
<tr>
<td>MRF-4</td>
<td>Normal</td>
<td>Myf-5 increases</td>
</tr>
</tbody>
</table>
**Figure 1 (A)** Type I functional divergence (Altered functional constraints) occurred after the gene duplication at an amino acid site between two duplicate genes. (B) A flowchart illustrating the statistical approach to test Type I functional divergence.
Figure 2. The NJ tree of Jaks, FGFRs and EGFRs based on the sequence alignment of kinase domains. The statistical reliability of the inferred tree topology was assessed by the bootstrap technique (Felsenstein 1985).
Figure 3. Functional divergence related amino acid residues candidate. (A) category I: conserved in tandem kinase domains (JH1), variable in pseudokinase domains (JH2); (B) category II: conserved in pseudokinase domains (JH2), variable in tandem kinase domains (JH1).
Figure 4. The neighbor-joining tree of MyoD gene family. The statistical reliability of tree topology is tested by bootstrapping with 1000 replicates.
**Figure 5 (A)** The myogenesis pathway regulated by MyoD gene family. 
**B** A star-like tree with Type I functional branch lengths $b_F$. 
**C** A proposed model of functional divergence in the evolution of MyoD gene family. "i" represents the function to receive inductive signals, "m" represents the function in myocyte differentiation. The boldfaced arrows represent some novel function developed after gene duplication.
CHAPTER III:
EVOLUTIONARY PATTERNS OF GENE FAMILIES
GENERATED IN THE EARLY STAGE OF VERTEBRATES

A paper published in the Journal of Molecular Evolution

Yufeng Wang and Xun Gu

Abstract

In this paper we have analyzed 49 vertebrate gene families that were generated in the early stage of vertebrates and/or shortly before the origin of vertebrates, each of which consists of three or four member genes. We have dated the first \(T_1\) and second \(T_2\) gene duplications of 26 gene families with three member genes. The means of \(T_1\) (594 mya) and \(T_2\) (488 mya) are largely consistent to a well-cited version of two-round (2R) genome duplication theory. Moreover, in most cases, the time interval between two successive gene duplications is large enough that the fate of duplicate genes generated by the first gene duplication was likely to be determined before the second one took place. However, the phylogenetic pattern of 23 gene families with four members is complicated;

---


2 Graduate student and Assistant professor, respectively, Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University.

3 Author for correspondence
only 5 of them are predicted by 2R model, but 11 families require an additional gene (or genome) duplication. For the rest (7 families), at least one gene duplication event had occurred before the divergence between vertebrate and Drosophila, indicating a possible conflict for the 4:1 rule (member gene ratio between vertebrates and invertebrates). Our results show that Ohno's 2R conjecture is valid as a working hypothesis for providing a most parsimonious explanation. Although for some gene families, additional gene duplication is needed, the credibility of the third genome duplication (3R) cannot be excluded.

**Key words:**
Gene (genome) duplication, vertebrate evolution, gene family, duplication time dating

**Introduction**

Gene duplication is believed to play an important role during evolution by providing opportunities to evolve new gene functions that can lead to novel morphologies and physiologies (Ohno 1970; Li 1983). Sequence accumulation by genome projects has shown a rapid increase of gene families in the early stage of vertebrates, which implies the existence of large-scale gene (genome) duplication(s) during this evolutionary period (Holland et al. 1994; Kasahara et al. 1996; Endo et al. 1997; Nadeau and Sankoff 1997; Ruddle 1997; Pebusque et al 1998; Suga et al. 1999). Since the ratio of member genes of human to fruitfly in these gene families is typically 2:1, 3:1 or 4:1, a two-round genome
duplication theory in the early lineage of vertebrates has become popular (e.g., Lundin

In spite of its simplicity and elegance (e.g., in the case of four mammalian Hox
gene clusters), recent studies have raised several questions about the classical two-round
genome duplication (2R) theory (Skrabanek and Wolfe 1998; Hughes 1999; Smith et al
1999). Their argument can be briefly described as Fig. 1B, where the 2R hypothesis
predicts a phylogenetic pattern ((A,B),(C,D)). This prediction does not always hold in
those gene families that were generated in the early stage of vertebrates (for a recent
review, see Skrabanek and Wolfe 1998). For example, if the phylogenetic pattern is
(((A,B), C), D), an additional gene duplication is needed so that the hypothesis of three-
round genome duplication (3R) is more favorable (Fig. 1C). However, testing 2R or 3R
theory is complicated by the fact that in many cases, two phylogenetic patterns,
((A,B),(C,D)) and ((A,B), C, D), are not statistically distinguishable or the topology is
sensitive to the tree-making method.

Some authors (e.g., Spring 1997) attempted to identify vertebrate gene families due
to 2R genome duplications, only using the synteny information and the member gene
ratio between vertebrates and invertebrates. Phylogenetic analysis of vertebrate gene
families has shown that such an approach can be misleading (Hughes 1999). However,
the view of Hughes (1999) that the role of genome duplication(s) in vertebrate evolution
has been disapproved seems to be extreme. Although the classical 2R model may be
oversimplified, the concept of genome duplication (Ohno 1970) in early vertebrates is
still valid as a working hypothesis. Otherwise, we are faced with the difficulty of
explaining the big-bang fashion of tissue-specific isoforms in the early stage of vertebrates. Indeed, in the human genome, a significant portion of gene families with two or three members is generated in the early vertebrate lineage, which are consistent with Ohno’s 2R hypothesis: one or more member genes may have been lost (Nadeau and Sankoff 1997). In this paper, we interpret the genome duplication hypothesis as only a few large-scale chromosome- or genome-wide duplication(s) can explain the generation of many duplicated genes during a relatively short time period in the early stage of vertebrate evolution. Although at this moment, it may be difficult to provide conclusive evidence for 2R, 3R or more round genome duplications (Skrabanek and Wolfe 1998), our understanding for vertebrate genome evolution can be significantly improved by exploring the evolutionary pattern of these gene families in a large scale.

Data and Methods

Amino acid sequences of vertebrate gene families were obtained from several databases including Genbank, EMBL and Hovergen (http://pbil.univ-lyon1.fr/). Drosophila and C. elegans homologues were obtained by BLAST-psi search. Two subsets of gene families were selected for our study.

Subset I: Gene families with three-gene clusters

From more than 1100 gene families, we have obtained 112 gene families with three gene clusters. Among them, 26 gene families meet the following criteria: (1) Each gene family has three member genes. (2) These member genes are likely to be generated
after the origin of vertebrates, which can be confirmed by using Drosophila and/or C. elegans homologues as an outgroup. (3) The Drosophila homologue can be reasonably considered as an orthologue of vertebrate gene clusters, so that the evolutionary distance between Drosophila and vertebrates can be used for calibration. (4) Amino acid sequences of each gene family should at least include species from mammals, birds or reptiles, and teleosts; and the phylogenetic tree should show a clear-cut evidence (i.e., bootstrapping value > 70%) that these duplicate genes were generated before the divergence between teleosts and tetrapods. And (5) Gene families with less than 100 amino acids (e.g. HOX) are not used because of statistical uncertainty. They are listed as follows (the number in the parenthesis is the number of sequences): Adrenergic Receptor α1, ADRA1 (16); Adrenergic Receptor α2, ADRA2 (29); Adrenergic Receptor β, ADRB (24); Aldolase, ALDO (18); Alzheimer β-Amyloid Precursor Protein, APLP (13); Calponin (19); Carboxypeptidase (14); Caudal, CDX (14); Cyclin D (15); Glycogen Phosphorylase, PYG (13); Early Growth Response Protein, EGR (12); Fibroblast growth factor 8/17/18, FGF 8/17/18 (12); Fos (18); GLI (10); Hedgehog, HH (20); Insulin Receptor, INSR (9); ITGA5/8/VNTR (10); Inositol 1,4,5-triphosphate Receptor, ITPR (10); JUN (16); MINOR (15); MYB (14); Nitric Oxidase Synthase, NOS (13); REL (14); ROR (14); Stimulating Hormone Receptor, SHR (26); Zonula Occludens, ZO (8).

Accession numbers of these sequences are available upon request.

*Subset II: Gene families with four-gene clusters*

In total, 58 gene families with four clusters (member genes) have been found. To elucidate the phylogenetic pattern, we selected 23 four-cluster gene families with
invertebrate homologues available. They are: Activin β (20); Cathepsin (20); Cyclic Nucleotide-gated Cation Channel, CNCG (18); Discs Large Protein, DLG (12); Epidermal Growth Factor Receptor, EGFR (13); ELAV (20); Ezrin (16); Glycine Receptor, GLR (12); ID (15); Janus Tyrosine Kinase, JAK (16); Myocyte Enhancer Factor 2, MEF2 (14); MAP Kinase Phosphatase, MKP (13); Myogenic Determining Factor, MyoD (30); NOTCH (14); Neuropeptide Receptor, NPYR (20); Octamer-binding Protein, OCT (23); Phosphodiesterase-4, PDE4 (10); 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase, PFKFB (15); Peroxidase, PO (12); RAS (23); Src-related (13); Syndecan (13); TRK (14).

It should be noted that the classification of phylogenetic patterns shown in Fig.1 could be complicated by the possibility that some member genes have not been sequenced yet. However, we think that it may not be the serious problem for our data. Indeed, new member gene search has been extensively conducted in many gene families since they can be easily found through the human/mouse EST data. Keeping this problem in mind, here we simply assume the effect of incompleteness is trivial. This assumption can be tested in the near future when the complete human genome sequence is available.

For each gene family, amino acid sequences were aligned by CLUSTALW (Thompson et al. 1994). The phylogenetic tree was reconstructed by the Neighbor-joining method (Saitou and Nei 1987). The root was determined by using the Drosophila and/or C.elegans homologous gene as an outgroup.

For each gene family with three-gene clusters, these two gene duplications were dated as follows. A linearized neighbor-joining tree (Takezaki et al. 1995) is used to
convert the (average) distance to the geological time scale when several speciation events are used for calibrations. We have fully recognized that it is highly controversial about the geological dates of vertebrates between paleontologists and molecular evolutionists (e.g., Kumar and Hedges 1998; Gu 1998). Therefore, it is necessary to test whether our results are sensitive to the choice of calibration points. Fortunately, our preliminary results have shown that the uncertainty of geological dates of vertebrates, i.e., primate-rodent \( t_1 = 65-115 \text{ mya} \), and tetrapod-teleost \( t_4 = 410-450 \text{ mya} \), has little effect on our time estimation of duplicate genes (data not shown). For simplicity, middle values (i.e. \( t_1 = 80 \text{ mya} \) and \( t_4 = 430 \text{ mya} \)) were used in our study. However, our estimation is indeed sensitive to the vertebrate-Drosophila divergence time \( t_5 \). Thus, two calibration systems were used to address this problem. The vertebrate calibrations are the same: primate-rodent \( t_1 = 80 \text{ mya} \), mammal-bird \( t_2 = 310 \text{ mya} \), mammal-amphibian \( t_3 = 350 \text{ mya} \), tetrapod-teleost \( t_4 = 430 \text{ mya} \), while the first calibration system uses the vertebrate-Drosophila \( t_5 = 830 \text{ mya} \) as an upper bound, according to the molecular data (Gu 1998). The second calibration time uses \( t_5 = 700 \text{ mya} \) as a lower bound, which seems reasonable because recent fossil records found in South China (Shu et al. 1999) imply that the first agnathans may have evolved in the earliest Cambrian (~ 555 mya). Our calibration systems minimize the effect of rate variation among vertebrate lineages since duplication events have occurred within the range of the calibration times, i.e., after the divergence time of vertebrate-Drosophila and definitely before that of primate-rodent. (Gu 1998).
Results

**Dating gene duplications in early vertebrates**

Fig. 2 shows the original and linearized phylogenetic tree for Aldolase (ALDO) gene family, which was inferred by the neighbor-joining method with Poisson distance. Bootstrap values more than 50% were presented. 1\textsuperscript{st} and 2\textsuperscript{nd} represent the time points (T\textsubscript{1} and T\textsubscript{2}, respectively) of the first and the second gene duplications, respectively. The linearized neighbor-joining tree was used to compute the average distance (d). We found that the linearity between d and the calibration time (t) is very good for both calibration systems I and II; the correlation coefficients for two calibration systems are 0.95 and 0.94, respectively. Thus, as shown in Fig. 1(C) and Fig. 1(D), we can estimate times of the first and second gene duplications.

The estimated times of first and second gene duplications in twenty-six gene families are presented in Table 1. At first we discuss the result based on the first calibration system (i.e., the split of vertebrate-Drosophila is about 830 mya). For most of cases (21 of 26), T\textsubscript{1} is larger than 500 mya, while for all except one case, T\textsubscript{1} is less than 700 mya. The mean of T\textsubscript{1} over all families is 594 mya, with standard error 17 mya. Similarly, the mean of T\textsubscript{2} over all families is 488 mya, with standard error 15 mya. It is noteworthy that the biological interpretation of these means of T\textsubscript{1} and T\textsubscript{2} is meaningful only under the 2R hypothesis, i.e., all these gene families were generated by two-round genome duplications in the early stage of vertebrates. At least for some of them, this assumption may not hold. This is because we cannot rule out the possibility that due to the incompleteness of data or losses of member genes, a gene duplication event can
actually be the result of a third genome duplication (3R model) or simply is irrelevant to any genome duplication. At any rate, this dating is largely consistent to the most widely cited version of the 2R hypothesis. (1) The first round of genome duplication is assumed to have occurred prior to the divergence of Agnatha, while the mean of $T_1$ is 594 mya. And (2) the second round of genome duplication is assumed to have occurred before the divergence of Chondryichthyes, while the mean of $T_2$ is 488 mya.

To explore the pattern of functional divergence after successive gene (genome) duplications, it is important to know the separation in geological time between them, i.e., $\delta = T_1 - T_2$ (Nadeau and Sankoff 1997). Two successive gene duplications may act as a single event in the time scale of long-term evolution if $\delta < 30$ million years (myr), which can be explained as follows. From the population genetics viewpoint, if two gene duplication events occurred in a very close time period, these (as many as four) duplicate genes are polymorphic that have not been fixed (or lost) in the population. As a result, the long-term effect may not be distinguished from one single event (Li 1983). As extensively discussed by Nadeau and Sankoff (1997), $\delta$ may provide biological insights for understanding the relative rate of functional divergence and loss after gene (genome) duplications. As shown in Table 1, the distribution of the separation of geological time ($\delta$) between $T_1$ and $T_2$ (the first calibration system) can be summarized as follows: four cases with $\delta < 50$ myr, nine cases for $50 < \delta < 100$ myr, and thirteen cases with $\delta > 100$ myr. The mean of $\delta$ is 106 myr with standard error 12 myr. That is, in most cases $\delta$ is significantly longer than the average survival time of a duplicate gene in the genome without functional divergence (Li 1983). It implies that the fate of most duplicate genes
from the first gene (genome) duplication might have been determined (either acquiring
some new functions or becoming a pseudogene) when the second gene (genome)
duplication took place. This result holds regardless of the debate whether these gene
families were generated by two or three round genome duplications, but has an indication
that these gene (genome) duplications may have distinct impacts on the evolution of
vertebrate genomes.

As expected, our estimates based on the second calibration system, on average,
are lower than those based on the first calibration system, though they are similar (Table
1). The mean of $T_1$ is 525 mya and that of $T_2$ is 435 mya. Since phylogenetic analysis of
each gene family clear show that these gene duplications had occurred before the
divergence between tetrapods and teleosts (i.e., >430 mya, see data and methods, results
not shown), it seems unreasonable if the estimate of $T_2$ is less than 400 mya. This can be
cased by the non-linear property of the calibration system, which is apparently non-
trivial in the second calibration system; indeed there are 10 such cases (Table 1). We
find that the split time between vertebrate and Drosophila ($t_s=700$ mya) is not consistent
with the vertebrate calibrations since it is not the case for the first calibration system.
Therefore, it seems that our results based on the first calibration system are more
reasonable. Interestingly, the distribution of the separation of geological time between $T_1$
and $T_2$ is fairly invariant, with a mean of 90 myr.

Instead of a Poisson distance, we have used various distances including PAM and
gamma distance to examine the robustness of our results; they usually give similar results
with ~5% margin (data not shown). To examine the extent to which our dating is
affected by the rate variation among lineages, we have calculated the relative time ratio of each gene duplication event to a single speciation event that has been used for calibration. For instance, the means of relative dates of the first gene duplication are 0.67, 1.78, and 2.71, respectively, to the divergence time of vertebrate-Drosophila, tetrapod-teleost, and mammal-bird; those of the second gene duplication are 0.48, 1.31 and 1.93, respectively. Therefore, if we only use vertebrate calibrations, the dates of these genome duplications can be inflated up to $T_1 \approx 800$ mya and $T_2 \approx 600$ mya, about 25% higher than our estimates even if the upper bound ($t_s = 830$ mya) is used. We believe that it may be caused by a fast evolution after gene duplication, which may be as a result of functional redundancy (Li 1983). If that is the case, $T_1 - T_2$ is also inflated. This effect has been significantly alleviated if Drosophila is used as an outgroup.

An important but difficult problem is how to deal with the gene clusters with different evolutionary rates. We indeed have observed that in several cases the evolutionary rates among clusters vary so much that could potentially affect the accuracy of time estimation. Fortunately, for most cases under study, the evolutionary rates among clusters do not vary significantly so that we could use the average rate as a compromise. We may take two approaches. The first one is our current approach, i.e., first compute the average evolutionary rate, and then compute the duplication time. The other approach is to compute the duplication time from each member gene, and then take an average. Though these two approaches are feasible in practice, we found that statistically the first one is more stable (data not shown).
Complicated patterns of four gene clusters

Phylogenetic analysis for 23 gene families with four member genes shows complicated patterns, as denoted by type I, II and III, respectively (Fig. 1). As summarized in Table 2, only five gene families show type I phylogenetic pattern that is predicted by the 2-R model. For example, the phylogenetic trees of Activin β and PFKF gene families (Fig. 3) show that four isoforms were likely to be generated after the divergence between vertebrates and Drosophila. However, 11 gene families show type II phylogenetic pattern (Fig. 1C). For instance, see Fig. 4 for ELAV (Hu) and RAS gene families. The branching points of duplication events are highly supported by the bootstrap test, indicating the statistical reliability of this tree topology (Table 2). Thus, the 2R model is not sufficient, and an additional gene (or genome) duplication event is needed. Moreover, the remaining seven gene families indicate that at least one gene duplication event is prior to the divergence between vertebrates and Drosophila. As an example, see Fig. 5 for Mef2 and MyoD gene families. Therefore, these four gene clusters should not be considered as the product of two-round genome duplication (Hughes 1999). Indeed, Mef2 gene family has been used as a “good” example according to the 4:1 rule (Spring 1997), which turns out to be incorrect. However, denying the generation of isoform Mef2B as the result of any possible genome duplication does not necessarily rule out the possibility that the origin of other three isoforms (Mef2A, D and C) are related to genome duplication(s).
Discussion

Is genome duplication hypothesis wrong?

Since the hypothesis of genome duplication, in particular the 2R model, has been widely cited in the field of vertebrate development and evolution, it deserves a rigorous test. Our exhaustive search over more than 1100 gene families have shown that 112 gene families with three members, 58 gene families with four members, and 130 gene families with two member genes; multiple families that are apparently generated recently such as Ig, TCR or Olfactory receptors have been excluded. Sequence diversity among member genes is usually >4 times larger than that between orthologous human/mouse genes, indicating that they have been maintained in the genome since the early stage of vertebrates. The appearance of such large number of multiple-cluster families (~27% in the vertebrate genome) implies the possibility of genome-wide duplication(s).

Further, we have analyzed 49 vertebrate gene families with at least three member genes to investigate the evolutionary pattern of gene (genome) duplications in the early stage of vertebrates or shortly before the origin of vertebrates. It has been clear that the two-round genome duplication theory (2R) can well explain the mean duplication times over 26 gene families with three gene clusters and the phylogenetic pattern (type I) of five gene families with four-gene clusters. However, to explain type II phylogenetic pattern of four gene clusters, 2R model may not be sufficient and a 3R model is needed, or we have to assume one of gene duplications is irrelevant to any genome-wide duplication. At this moment, we cannot rule out any possibility. Seven cases in type III phylogenetic pattern shows that it can be very misleading to assign any four (or more)
tissue-specific genes in the human genome as “potential” products of the 2R genome duplication, without any detailed phylogenetic analysis. For example, we have already found several such mistakes in the Tetrabase developed by Spring (1997).

However, we do not agree Hughes (1999)'s final conclusion that the role of genome duplication in vertebrate evolution is entirely speculative. What we have shown in this paper, as well as many authors (Skrabanek and Wolfe 1998; Hughes 1999; Smith 1999), is that the generation of gene families in the early vertebrate is much more complicated than we previously thought. In principle, there is a simple correlation between genome duplication(s), chromosome synteny, phylogeny, and functional divergence (e.g., tissue-specificity). It has been shown that current vertebrate genomes contain substantial such information (e.g., Lundin 1993; Holland et al. 1994; Kasahara et al. 1996; Endo et al 1997; Nadeau and Sankoff 1997; Ruddle 1997; Pebusque et al 1998; Suga et al. 1999). Although this type of information is invaluable for understanding the origin of human genome complexity, the difficulty for a direct testing by using these data is that in the past several hundred million years, the vertebrate genome has experienced dramatically chromosome rearrangement, functional divergence and gene loss. Without a stochastic model that takes these mechanisms into account, it is not surprising that our attempt to prove genome duplication(s) is always unsuccessful (Skrabanek and Wolfe 1998). Nevertheless, the 2R model clearly provides a most parsimonious explanation for most observations; an addition gene duplication for some gene families under study is necessary but it remains unclear whether its scale is large enough to be considered as a third genome duplication (the 3R) model. If the complete sequence of human genome
finally shows a significant portion of duplicate genes is indeed generated during a short
time period of the early vertebrate lineage, it is difficult to imagine how can we offer an
explanation without the concept of genome duplication. At any rate, adding an additional
(or even more) genome duplication to the conventional 2R model is fundamentally
different from denying the concept of genome duplication. Currently we are working on
these challenging but exciting problems.

Acknowledgements

This study was supported by an ISU startup fund and an NSF grant to X. G. We
thank Jianying Gu for assistance. We are grateful to an anonymous referee for his helpful
comments.

References

Endo T, Imanishi T, Gojobori T, Inoko H (1997) Evolutionary significance of intra-
genome duplications on human chromosomes. Gene 205:19-27
47: 369-371
Holland PWH, Garcia-Fernandez J, Willaime NA, Sidow A. (1994) Gene duplication and
the origins of vertebrate development. Development supplement 125-133
Hughes AL (1999) Phylogeny of developmentally important proteins do not support the
hypothesis of two rounds of genome duplication early in vertebrate history. J Mol
Evol 48:565-76


Spring J (1997) Vertebrate evolution by interspecific hybridization - are we polyploid? FEBS Letters 400:2-8


**Table 1**: Time estimation of the first and the second gene duplications of 26 gene families with three-gene clusters.

<table>
<thead>
<tr>
<th>Gene Family</th>
<th>Sites</th>
<th>Calibration System I</th>
<th>Calibration System II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$T_1$(mya)</td>
<td>$T_2$(mya)</td>
</tr>
<tr>
<td>EGR</td>
<td>308</td>
<td>724</td>
<td>624</td>
</tr>
<tr>
<td>Aldo</td>
<td>360</td>
<td>699</td>
<td>404</td>
</tr>
<tr>
<td>Myb</td>
<td>506</td>
<td>691</td>
<td>528</td>
</tr>
<tr>
<td>Nos</td>
<td>1100</td>
<td>690</td>
<td>563</td>
</tr>
<tr>
<td>ZO</td>
<td>821</td>
<td>685</td>
<td>458</td>
</tr>
<tr>
<td>Carboxy-peptidase</td>
<td>390</td>
<td>682</td>
<td>598</td>
</tr>
<tr>
<td>APLP</td>
<td>501</td>
<td>657</td>
<td>473</td>
</tr>
<tr>
<td>ROR</td>
<td>431</td>
<td>643</td>
<td>444</td>
</tr>
<tr>
<td>ITPR</td>
<td>2561</td>
<td>632</td>
<td>614</td>
</tr>
<tr>
<td>Rel</td>
<td>380</td>
<td>629</td>
<td>461</td>
</tr>
<tr>
<td>ITGA5/8/V</td>
<td>981</td>
<td>628</td>
<td>619</td>
</tr>
<tr>
<td>NTR</td>
<td>1199</td>
<td>623</td>
<td>589</td>
</tr>
<tr>
<td>Insulin Receptor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hedgehog</td>
<td>277</td>
<td>618</td>
<td>552</td>
</tr>
<tr>
<td>Caudal</td>
<td>145</td>
<td>616</td>
<td>495</td>
</tr>
<tr>
<td>Jun</td>
<td>239</td>
<td>598</td>
<td>471</td>
</tr>
<tr>
<td>Pyg</td>
<td>836</td>
<td>595</td>
<td>452</td>
</tr>
<tr>
<td>ADRA1</td>
<td>383</td>
<td>592</td>
<td>480</td>
</tr>
<tr>
<td>MINOR</td>
<td>276</td>
<td>579</td>
<td>496</td>
</tr>
<tr>
<td>SHR</td>
<td>473</td>
<td>568</td>
<td>515</td>
</tr>
<tr>
<td>Cyclin D</td>
<td>279</td>
<td>541</td>
<td>468</td>
</tr>
<tr>
<td>GLI</td>
<td>392</td>
<td>520</td>
<td>413</td>
</tr>
<tr>
<td>FGF 8/17/18</td>
<td>189</td>
<td>490</td>
<td>423</td>
</tr>
<tr>
<td>Fos</td>
<td>242</td>
<td>472</td>
<td>410</td>
</tr>
<tr>
<td>Calponin</td>
<td>174</td>
<td>440</td>
<td>377</td>
</tr>
<tr>
<td>ADRA2</td>
<td>287</td>
<td>424</td>
<td>400</td>
</tr>
<tr>
<td>ADRB</td>
<td>310</td>
<td>419</td>
<td>362</td>
</tr>
</tbody>
</table>

Note: All estimates are based on the calibration time of primate-rat (80 mya); mammal-bird (310 mya), mammal-amphibian (350 mya), tetrapod-teleost (430 mya), plus different calibration times for the vertebrate-Drosophila split. Calibration system 1 is based on the molecular dating of vertebrate-Drosophila split (830 mya) (Gu 1998), whereas calibration system 2 assumes the divergence time of vertebrate-Drosophila as 700 mya. Note that the controversy about the time of mammal-mammal split (65-115 mya) (Kumar and Hedges 1998) has almost no effect on our estimation. (mya: million years ago; myr: million years).
Table 2. Phylogenetic patterns of four-gene cluster families that have been used in this study.

<table>
<thead>
<tr>
<th>Tree Topology</th>
<th>Gene Family</th>
<th>Aligned Sites</th>
<th>Bootstrap value(s)(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type I</strong> (((A,B),(C,D)),invertebrate)</td>
<td>Activin β</td>
<td>293</td>
<td>85(^a), 99(^b)</td>
</tr>
<tr>
<td></td>
<td>EGFR</td>
<td>634</td>
<td>97; 99</td>
</tr>
<tr>
<td></td>
<td>ID</td>
<td>97</td>
<td>96; 57</td>
</tr>
<tr>
<td></td>
<td>Jak</td>
<td>663</td>
<td>99; 98</td>
</tr>
<tr>
<td></td>
<td>PFKFB</td>
<td>388</td>
<td>97; 83</td>
</tr>
<tr>
<td></td>
<td>CNCG</td>
<td>537</td>
<td>99(^c); 69(^d)</td>
</tr>
<tr>
<td></td>
<td>DLG</td>
<td>679</td>
<td>33; 99</td>
</tr>
<tr>
<td></td>
<td>ELAV</td>
<td>243</td>
<td>98; 95</td>
</tr>
<tr>
<td></td>
<td>GLR</td>
<td>328</td>
<td>99; 99</td>
</tr>
<tr>
<td></td>
<td>MKP</td>
<td>254</td>
<td>96; 76</td>
</tr>
<tr>
<td></td>
<td>PDE4</td>
<td>435</td>
<td>99; 90</td>
</tr>
<tr>
<td></td>
<td>PO</td>
<td>687</td>
<td>99; 99</td>
</tr>
<tr>
<td></td>
<td>Ras</td>
<td>179</td>
<td>76; 76</td>
</tr>
<tr>
<td></td>
<td>Src-related</td>
<td>452</td>
<td>99; 99</td>
</tr>
<tr>
<td></td>
<td>Syndecan</td>
<td>156</td>
<td>88; 92</td>
</tr>
<tr>
<td></td>
<td>Trk</td>
<td>711</td>
<td>99; 99</td>
</tr>
<tr>
<td><strong>Type II</strong> (((A,B),(C),D),invertebrate)</td>
<td>Cathepsin</td>
<td>312</td>
<td>98(^e); 99(^f)</td>
</tr>
<tr>
<td></td>
<td>Ezrin</td>
<td>552</td>
<td>99; 99</td>
</tr>
<tr>
<td></td>
<td>Mef2</td>
<td>291</td>
<td>99; 99</td>
</tr>
<tr>
<td></td>
<td>Notch</td>
<td>1052</td>
<td>96; 99</td>
</tr>
<tr>
<td></td>
<td>NPYR</td>
<td>249</td>
<td>92; 99</td>
</tr>
<tr>
<td><strong>Type III</strong> (((A,B),(C),invertebrate),D)</td>
<td>MyoD</td>
<td>185</td>
<td>99(^g); 99(^h)</td>
</tr>
<tr>
<td>(((A,B),(invertebrate),(C,D))</td>
<td>OCT</td>
<td>219</td>
<td>99; 99</td>
</tr>
</tbody>
</table>

Note: See Fig. 1 for the topologies of three types of phylogenetic pattern. Invertebrate (Drosophila Melanogaster or and C. elegans) were used as an outgroup.

\(^{a,b}\): Bootstrap values of branching points D2 in Fig 1B for Type I gene family.

\(^{c,d}\): Bootstrap values of branching points D2 and D3 in Fig 1C for Type II gene family, respectively. Only for one case (DLG gene family), the Bootstrap value = 35\%, which should be considered as ambiguous.

\(^{e-h}\): Bootstrap values of branching points Ds in Fig 1D for Type III gene family.
Fig. 1. Phylogenies of gene families which have undergone gene duplications in the early vertebrate stage. Drosophila melanogaster or/and C. elegans were used as an invertebrate outgroup. (A) A topology of gene families having three gene clusters which were likely generated by two gene (genome) duplications (D1 and D2, respectively). Panels (B, C, D) show three possible phylogenetic patterns of gene families with four member genes. (B) Type I topology that has been possibly arisen from two rounds of gene duplications. (C) Type II topology that has been possibly arisen from three round genome duplications (D1, D2, and D3). (D) Type III topology in which at least one gene duplication had occurred before the divergence of invertebrates and vertebrates.
Fig. 2. Phylogenetic analysis of aldolase gene family. (A) The original phylogenetic tree. Neighbor-joining method with Poisson distance was used to infer the tree topology. Bootstrapping values more than 50% were presented. 1st and 2nd represent the time points of the first and the second gene duplications, respectively. (B) The linearized neighbor-joining tree which was used to compute the average distance. Bootstrapping values more than 50% were presented. (C) The linear relationship of evolutionary distance (d) against time (t) (correlation coefficient r = 0.95) based on the Calibration System I. (D) The linear relationship of evolutionary distance (d) against time (t) (correlation coefficient r = 0.94) based on the Calibration System II.
Fig. 3. Examples of N-J trees for Type I gene families with four member genes. (A) Activin b gene family. (B) PFKF gene family.
Fig. 4. Examples of N-J trees for Type II gene families with four member genes. (A) Elav (Hu) gene family. (B) Ras gene family.
Fig. 5. Examples of N-J trees for Type III gene families with four member genes. (A) Mef-2 gene family. (B) Ras gene family.
CHAPTER IV:
PATTERNS OF FUNCTIONAL DIVERGENCE AFTER
GENE DUPLICATIONS IN VERTEBRATE GENE
FAMILIES: ALTERED FUNCTIONAL CONSTRAINTS

A paper published in the Proceedings of the Atlantic Symposium on Computational Biology, Genome Information Systems & Technology.¹

Yufeng Wang²,³, Jianying Gu² and Xun Gu²

Introduction

Gene duplication and subsequent functional divergence are believed to give rise to evolutionally related but functional distinct paralogous genes in a variety of vertebrate genomes [1]. After gene duplication, one gene copy keeps the original function under the evolutionary constraint, while the other copy is redundant and free to accumulate mutations, which may eventually become a pseudogene or acquire a novel function [2]. Several alternative models have been proposed recently [3,4]. Although the detail of

² Graduate students and Assistant Professor, respectively, Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University.
³ Primary researcher and author.
functional divergence after gene duplication still remains largely unknown, altered functional constraints between homologous member genes of a family seems to be a good indicator for functional divergence [5-7]. Based on the principle that the changes of functional importance is highly correlated with the change of evolutionary conservation, Gu [7] developed a probabilistic model that can distinguish between the amino acid changes leading to the functional divergence and the background “noise” representing neutral evolution. In this paper, our purpose is to understand the impacts of gene duplications on the functional divergence in vertebrate gene families. Two patterns of functional divergence after gene duplication(s) have been found, indicating two rounds of gene duplications [8] may have distinct roles in the functional diversification.

Methods

The data set

Complicated evolutionary patterns of gene families generated in the early vertebrate stage have been discussed in Wang and Gu [9]. A subset of gene families with three gene clusters is used in this study (Table 1). The phylogenetic analysis as well as chromosome syntenies shows that these 10 gene families may have undergone two rounds (2R) of gene duplications after the emergence of early vertebrates (Fig. 1A).
Statistical measure of Type I functional divergence

The *type I functional divergence* refers to a particular type of functional divergence after gene duplication that results in altered functional constraints (i.e., different evolutionary rates) between two duplicate genes [7].

Under a two-state probabilistic model, an amino acid site has two possible states in each gene cluster, a functional divergence-unrelated state \((F_0)\) or a functional divergence-related state \((F_1)\). In combination, there are four possibilities for the gene family with two member genes A and B, \(F_0, F_0\), \(F_0, F_1\), \(F_1, F_0\), and \(F_1, F_1\). \(F_0, F_0\) means no altered functional constraints in both gene clusters, i.e., the evolutionary rate is virtually the same between two clusters. For the other combinations, however, the amino acid residue has experienced altered functional constraints at least in one cluster, resulting in rate independence between gene clusters. Thus, we have two (nondegenerate) combined states, i.e. \(S_0=(F_0, F_0)\) and \(S_1=(F_0, F_1) \cup (F_1, F_0) \cup (F_1, F_1)\). \(S_0\) and \(S_1\) are called *functional divergence configuration* of the gene family (Fig. 1B).

The *coefficient of type I functional divergence* \((\theta)\) between two gene clusters is defined as the probability of a site being functional divergence-related at least in one gene cluster, i.e., \(\theta = P(S_1)=1-P(S_0)\). A fast algorithm was developed for estimating \(\theta\) from a multiple alignment of amino acid sequence when the phylogeny of the gene family is given [7]. If \(\theta\) is significantly greater than 0, it provides evidence that altered functional constraints may occur after gene duplication.
Mapping functional divergence to a specific subfamily

For a gene family with multiple \((n)\) gene clusters, we can estimate the coefficient of (type I) functional divergence for each pair of gene clusters. Here we develop a simple method to show the general pattern of functional divergence of the gene family. By definition, the coefficient \((\theta_{AB})\) of type I functional divergence between clusters \(A\) and \(B\) follows the relation \(1-\theta_{AB} = P(S_0^{[AB]} = P(F_0^{[A]}, F_0^{[B]}).\) Let \(\theta_A = P(F_1^{[A]}\) and \(\theta_B = P(F_1^{[B]}\) be the probabilities of a site being functional divergence-related in clusters \(A\) and \(B\), respectively. Therefore, under the assumption of statistical independence of type I functional divergence between clusters, i.e., \(P(F_0^{[A]}, F_0^{[B]} = P(F_0^{[A]} \times P(F_0^{[B]}),\) we have

\[
1-\theta_{AB} = (1-\theta_A)(1-\theta_B). \tag{1}
\]

For convenience, a simple distance measure for functional divergence can be defined as

\[
d_F(AB) = -\ln (1-\theta_{AB})
\]

\[
b_F(A) = -\ln (1-\theta_A), \quad b_F(B) = -\ln (1-\theta_B) \quad \tag{2}
\]

From Eqs (1) and (2), it is easy to show that the measure \(d_F(AB)\) is additive, i.e.

\[
d_F(AB) = b_F(A) + b_F(B). \tag{3}
\]

Therefore, \(d_F(AB)\) is called \textit{type I functional distance} between clusters \(A\) and \(B\), and \(b_F(A)\) \([or \, b_F(B)]\) is called \textit{type I functional branch length} of clusters \(A\) \([or \, B]\), respectively.

When the coefficient of type I functional divergence for each pair of clusters is estimated by Gu [7], the matrix of type I functional distances can be easily obtained by Eq.(2). A standard least square method is implemented based on Eq.(3) to estimate the type I functional branch length for each gene cluster. This approach is used to illustrate the pattern of type I functional divergence among member genes in a multi-gene family.
In the case of three-cluster gene families, the least square solutions for functional branch lengths are given by

\[
\begin{align*}
    b_{F}(1) &= \frac{d_{F}(12) + d_{F}(13) - d_{F}(23)}{2}; \\
    b_{F}(2) &= \frac{d_{F}(12) + d_{F}(23) - d_{F}(13)}{2}; \\
    b_{F}(3) &= \frac{d_{F}(13) + d_{F}(23) - d_{F}(12)}{2},
\end{align*}
\]

respectively.

**Results and Discussion**

**Coefficient of type I functional divergence**

Table 2 shows that the pair-wise coefficients (\(\theta\)) of type I functional divergence between three member genes of 10 gene families. \(\theta\) value ranges from 0.001 to 0.915. Among 30 pair-wise \(\theta\) values, 28 are significantly greater than 0, indicating that the altered functional constraints are likely to be common after gene duplication(s).

**Two patterns of functional divergence in 3-cluster gene families**

We obtained functional distances (\(d_{F}\)) and functional branching lengths (\(b_{F}\)) for each gene cluster based on \(\theta\) values [Eqs.(2) and (4)]. In the functional distance analysis, in principle, virtually zero functional branch length of a gene cluster indicates the ancestral function, while a long branch length indicates diversified function. Two major patterns of functional divergence after two rounds of gene duplication have been found. For convenience, we denote Clusters 1 and 2 as one clade after two rounds of gene
duplication, whereas the other clade only includes Cluster 3, with the possibility that the other gene copy within this clade is not found or extinct in the evolution (Fig 1).

**Pattern I: one copy after the 1st round of gene duplication maintains the original function**

Three gene families (ALDO, HH, NOS) share the common pattern: Cluster 3 shows virtually zero b_F values (Table 3). Although we cannot rule out the possibility that the not-yet-found or lost paralog of Cluster 3 may have distinct function, it is reasonable to speculate that after the 1st round of gene duplication, one of the copies (the ancestor of Cluster 3) preserves the ancestral function. This may have freed the other copy (the ancestor of Clusters 1 and 2) from selective constraint, allowing it to diverge and take on new roles in specific tissues or organs. Interestingly, these three families show two different sub-patterns of functional divergence after the 2nd round of gene duplication:

**Sub-pattern I**

In ALDO gene family, Clusters 1 (AldoA) and 2 (AldoC) both have long functional branch lengths (0.489 and 0.364, respectively). Experimental evidence has shown that AldoA and AldoC are tissue-specific: AldoA is only expressed in the fibroblast, while AldoC only functions in the brain [10]; Similarly, In NOS gene family, Clusters 1 (NOS1) and 2 (E-NOS) show considerably large b_F values (0.431 and 1.302, respectively) (Fig. 2A), which is consistent with their strong tissue specificity: NOS1 is expressed in neural system and uniquely participates neurotransmission, and E-NOS is specifically expressed in endothelial tissues [11]. In contrast, I-NOS, which has virtually zero (-.085) functional branch length, is found generally expressed in several tissues such
as heart and muscles. One may speculate that the tissue-specific functions of Clusters 1 and 2 in ALDO and NOS gene families may be developed under the shield of Cluster 3 which receives most selection pressure.

**Sub-pattern II**

In HH gene family, not only Cluster 3 (DHH) but also Cluster 1 (SHH) shows virtually zero $b_F$ value, while the $b_F$ value (0.117) of Cluster 2 (IHH) is significantly greater than 0. This is consistent with the result of functional assay: vertebrate SHH genes typically show conserved expression in the notochord and floor plate, while IHH shows different expression patterns in different tissues in vertebrates [12]. After the 2nd round of gene duplication, one of the copies (SHH, in this case) had to maintain the ancestral function, whereas IHH might be the one that acquired some new functions and had escaped from the purifying selection.

Indeed, pattern I suggests that the 1st round of gene duplication plays a role in the functional divergence by providing the opportunity of a duplicated copy to evolve new function.

**Pattern II: one copy after the 2nd round of gene duplication maintains the original function**

In contrast to the three gene families showing Pattern I, the rest seven gene families show Pattern II: the distance branch length of Cluster 3 is significantly greater than 0, indicating that Cluster 3 may represent some novel function after two rounds of
gene duplication; moreover, one gene copy after the 2\textsuperscript{nd} round of gene duplication (Cluster 1 or Cluster 2) shows virtually zero \( b_F \) value, whereas the other copy (Cluster 2 or Cluster 1) has long functional branch length. Extensive evidence has shown that certain level of functional divergence had occurred at least after the 2\textsuperscript{nd} round of gene duplications, although whether the 1\textsuperscript{st} round had detectable impacts on the functional divergence remains unclear.

For example, in the myb gene family, dramatic functional innovations are likely to occur after both 1\textsuperscript{st} and 2\textsuperscript{nd} rounds of gene duplications. Myb gene family consists of three member genes A, B, and c-myb that encode nuclear proteins functioning as transcription activators [13]. The 1\textsuperscript{st} round of gene duplication gave rise to B-myb, whereas the 2\textsuperscript{nd} produced A-myb and c-myb. Interestingly, A-myb and c-myb exert a negative regulatory effect on their transcription activation function, while B-myb functions as a positive regulator. This evidence supports the result of our functional distance analysis (\( b_F = 0.478 \) for B-Myb), implying that the shift of +/- regulation may have taken place after the 1\textsuperscript{st} round of gene duplication (Fig. 2B). Moreover, gene knockout experiments show that A-myb and c-myb are functionally distinct as well. Loss of c-myb function results in embryonic lethality. In contrast, A-myb null mice are viable but exhibit growth abnormalities. This result is consistent with the inference from our functional distance analysis. A-myb with \( b_F \) value 0.034 may represent the ancestral function, whereas c-myb (\( b_F = 0.740 \)) may carry on the novel function that is indispensable in development. Intuitively, these altered functional constraints reflect c-myb may be temporarily escaped from the evolutionary pressure by the presence of A-
myb. Hence, the functional evolution of myb member genes indicate that the 1st round of gene duplication generated two regulation types (+/-), and the 2nd gave rise to the distinct level of negative regulation.

Analogous to the case of Myb gene family, the pattern of +/- regulation changes after gene duplication is also observed in the Jun gene family. Jun gene family encodes a component of the transcription factor Ap-1 (fos/Jun) which plays multiple roles in functional development of hematopoietic cells and regulation of apoptosis (programmed cell death). The large b_F value (1.402) of Cluster 3 (Jun-B) implies along the lineage leading to it (i.e., after the 1st round of gene duplication) there might be substantial functional innovation. This implication is supported by the experimental evidence: Jun-B exhibits strong negative regulation of apoptosis, whereas Jun-D and c-jun both regulate apoptosis positively. The altered functional constraints also appear in Clusters 1 (b_F = -0.394 for Jun-D) and Cluster 2 (b_F = 1.063 for c-Jun). Indeed, functional assay shows that in contrast to Jun-D which only employs positive regulation, C-Jun has dual functional roles: positive regulation of apoptosis when it is solely present, strong negative role with Jun-B [14]. One may speculate that the 1st round of gene duplication may have contribution to the differentiation of +/- negative control of apoptosis, whereas the 2nd round may play a role in the fine regulation.

Another example is the SHR (Stimulating hormone receptor) gene family. Although the detailed functional role of Cluster 3 (FSHR) is still unclear, the expression patterns of Clusters 1 (LSHR) and 2 (TSHR) may provide an insight into the underlying mechanism of the functional divergence after the 2nd of gene duplication. LSHR has b_F =
-0.066 which indicates its low level functional divergence, while TSHR shows $b_F = 0.386$, which indicates high level of functional diversification. This significant difference in the functional branch length of LSHR and TSHR may reflect their tissue specificities: in contrast to LSHR which is widely expressed in both gonadal cells and thyroid, TSHR is only uniquely expressed in thyroid. The sequence analysis suggests that tissue specific splicing have occurred after the 2\textsuperscript{nd} round of gene duplication [15].

In summary, in Pattern II, Cluster 3 shows significant level of functional divergence after the 1\textsuperscript{st} round of gene duplication; and moreover, the two gene copies after gene duplications exhibit altered functional constraints, one of which seems to have inherited the ancestral function, and the other has developed new function.

**Conclusions**

We have performed a new statistical test (functional distance analysis) on ten 3-cluster vertebrate gene families that have been generated after two rounds of gene duplications. The major conclusions drawn from this study are: (1) Altered functional constraints is a common pattern in the vertebrate gene families after gene duplication(s); (2) Two patterns of functional divergence have been found. Both patterns show that at least of one copy of duplicated gene has virtually zero functional branch length ($b_F$), indicating it preserves ancestral function. In theory, this gene copy provides a buffering system for the other copy to evolve new functions with no or low evolutionary pressure. (3) Functional distance analysis may provide a simple measure for the level of functional
divergence between gene clusters after gene duplication(s) and further shed light on the
mechanism of functional innovations in functional genomics.

Acknowledgements

This study was supported by NIH grant ROI GM62118 to X.G.

References


Table 1. Three-cluster vertebrate gene families used in this study.

<table>
<thead>
<tr>
<th>Gene family</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADRA1 (Adrenergic receptor α1)</td>
<td>ADRA1A</td>
<td>ADRA1B</td>
<td>ADRA1C</td>
</tr>
<tr>
<td>ADRA2 (Adrenergic receptor α2)</td>
<td>ADRA2A</td>
<td>ADRA2C</td>
<td>ADRA2B</td>
</tr>
<tr>
<td>ADRB (Adrenergic receptor β)</td>
<td>ADRB1</td>
<td>ADRB2</td>
<td>ADRB3</td>
</tr>
<tr>
<td>ALDO (Aldolase)</td>
<td>AldoA</td>
<td>AldoC</td>
<td>AldoB</td>
</tr>
<tr>
<td>CDX (Caudal)</td>
<td>CDX1</td>
<td>CDX2</td>
<td>CDX4</td>
</tr>
<tr>
<td>HH (Hedgehog)</td>
<td>SHH</td>
<td>IHH</td>
<td>DHH</td>
</tr>
<tr>
<td>Jun</td>
<td>c-Jun</td>
<td>Jun-D</td>
<td>Jun-B</td>
</tr>
<tr>
<td>Myb</td>
<td>c-myb</td>
<td>A-myb</td>
<td>B-myb</td>
</tr>
<tr>
<td>NOS (Nitric oxidase synthase)</td>
<td>NOS1</td>
<td>E-NOS</td>
<td>I-NOS</td>
</tr>
<tr>
<td>SHR (Stimulating hormone receptor)</td>
<td>LSHR</td>
<td>TSHR</td>
<td>FSHR</td>
</tr>
</tbody>
</table>
Table 2. The coefficients of type I functional divergence of pair-wise comparison of three member genes.

<table>
<thead>
<tr>
<th>Gene family</th>
<th>( \theta_{AB} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cluster 1 vs. 2</td>
</tr>
<tr>
<td>ADRA1 (Adrenergic receptor ( \alpha_1 ))</td>
<td>0.746(0.167)</td>
</tr>
<tr>
<td>ADRA2 (Adrenergic receptor ( \alpha_2 ))</td>
<td>0.447(0.086)</td>
</tr>
<tr>
<td>ADRB (Adrenergic receptor ( \beta ))</td>
<td>0.161(0.095)</td>
</tr>
<tr>
<td>ALDO (Aldolase)</td>
<td>0.574(0.093)</td>
</tr>
<tr>
<td>CDX (Caudal)</td>
<td>0.004(0.094)</td>
</tr>
<tr>
<td>HH (Hedgehog)</td>
<td>0.099(0.044)</td>
</tr>
<tr>
<td>Jun</td>
<td>0.488(0.135)</td>
</tr>
<tr>
<td>Myb</td>
<td>0.595(0.073)</td>
</tr>
<tr>
<td>NOS (Nitric oxidase synthase)</td>
<td>0.823(0.126)</td>
</tr>
<tr>
<td>SHR (Stimulating hormone receptor)</td>
<td>0.274(0.102)</td>
</tr>
</tbody>
</table>

Note: the numbers in the parentheses are standard errors of \( \theta \) values.
Table 3. The functional branch lengths ($b_F$) of three gene clusters.

<table>
<thead>
<tr>
<th>Gene family</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALDO</td>
<td>0.489</td>
<td>0.364</td>
<td>-0.22</td>
</tr>
<tr>
<td>NOS</td>
<td>0.431</td>
<td>1.302</td>
<td>-0.085</td>
</tr>
<tr>
<td>HH</td>
<td>-0.0125</td>
<td>0.117</td>
<td>0.072</td>
</tr>
<tr>
<td>ADRA1</td>
<td>-0.01</td>
<td>1.38</td>
<td>0.25</td>
</tr>
<tr>
<td>ADRA2</td>
<td>0.533</td>
<td>0.056</td>
<td>0.267</td>
</tr>
<tr>
<td>ADRB</td>
<td>-0.172</td>
<td>0.348</td>
<td>0.731</td>
</tr>
<tr>
<td>CDX</td>
<td>-0.118</td>
<td>0.122</td>
<td>0.119</td>
</tr>
<tr>
<td>Jun</td>
<td>1.063</td>
<td>-0.394</td>
<td>1.402</td>
</tr>
<tr>
<td>Myb</td>
<td>0.74</td>
<td>0.034</td>
<td>0.478</td>
</tr>
<tr>
<td>SHR</td>
<td>-0.066</td>
<td>0.386</td>
<td>0.609</td>
</tr>
</tbody>
</table>
Invertebrate outgroup

Figure 1. Phylogeny of three-cluster gene families that have undergone two rounds of gene duplications
Figure 2. (A) The phylogenetic tree topology of NOS gene family and a star-like tree scaled with functional branch lengths which shows Pattern I functional divergence. (B) The phylogenetic tree topology of Myb gene family and a star-like tree scaled with functional branch lengths that shows Pattern II functional divergence.
GENERAL CONCLUSIONS

This study focuses on a statistical and phylogenetic approach to investigate the functional divergence and evolutionary patterns among gene families. The author believes that a combination of hierarchical information (sequence, structural, functional, expression, etc. information), and cutting edge technologies (computational and experimental) will greatly enhance our understanding of functional genome innovation.

Summary of Findings

In the initial part of this study, we investigated the functional divergence (site-specific rate difference) in the caspase gene family that is crucial for programmed cell death and cytokine processing. Our results are summarized as follows. (1) Phylogenetic analysis shows that the major caspase-mediated pathways have been coordinately evolved after gene duplications. (2) Type I functional divergence (altered functional constraints) is statistically significant between two major subfamilies, CED-3 and ICE. (3) 21 sites are predicted to be mainly responsible for the functional divergence between two subfamilies. And the critical roles of four predicted amino acid residues have been verified by experimental evidence. (4) Some CED-3 caspases may inherit a large component of ancestral functions, while others may carry on more recently-derived functions. This study shows the potential of making cost-effective prediction from statistical modeling on sequence evolution (Gu 1999).

In Chapter 2, we describe the framework of the new statistical method phylogeny-based analysis (PHABA) (Gu 1999) and then provide many examples to show the
association between computational predictions and other biological information, including three-dimension protein structure, knock-out mice and mutagenesis evidence. Three components of PHABA have important biological inferences: (1) the coefficient of functional divergence ($\Theta$) is an index for the statistical significance of the level of functional divergence; (2) A site-specific profile of probability being functional divergence related provides a cost-effective measure prior to experimentation. (3) The functional distance is a quantitative measure for the contribution of each gene cluster to the overall functional divergence, i.e., cluster-specific functional divergence. Our results show PHABA can make statistically sound and biologically meaningful predictions from sequences and thus will benefit functional and evolutionary genomics.

In Chapter 3, we surveyed 49 vertebrate gene families arisen in the early stage of vertebrates and/or shortly before the origin of vertebrates, each of which consists of three or four member genes. The times of the first ($T_1$) and second ($T_2$) gene duplications in 26 three-cluster gene families are estimated to be 594 mya and 488 mya, respectively. This result essentially agrees with the well-recognized two-round (2R) genome duplication theory. Moreover, in most cases, the time interval between two successive gene duplications is large enough that the fate of duplicate gene copies generated by the first gene duplication was likely to be determined before the second one occurred. However, the evolutionary pattern of 23 gene families with four members is complicated; only 5 of them are predicted by 2R model, 11 families require an additional gene duplication. In the remaining 7 families, at least one gene duplication event had occurred before the divergence between vertebrate and Drosophila, indicating a possible misleading of the
4:1 rule (member gene ratio between vertebrates and invertebrates). Thus, Ohno's 2R conjecture is still valid as a working hypothesis for providing a most parsimonious explanation, although additional gene duplication is needed for some gene families, and the credibility of the third genome duplication (3R) remains unclear.

After methodology validation in Chapters 1 and 2, in Chapter 4, we performed a new statistical test (functional distance analysis) on ten 3-cluster vertebrate gene families generated after two rounds of gene duplications (data sets from Chapter 3). The major findings are: (1) Altered functional constraints is a general pattern in the evolution of vertebrate gene families after gene duplication(s); (2) Two patterns of functional divergence have been found. Both patterns show that at least one copy of duplicated gene has virtually zero functional branch length ($b_f$), indicating an ancestral function. In theory, this gene copy provides a buffering system for the other copy to evolve new functions with no or low evolutionary pressure. (3) Functional distance analysis may provide a simple measure for the level of functional divergence between gene clusters after gene duplication(s).

**Ongoing and Future Work**

With the promising results obtained from the theoretical study and testing in the functional divergence and evolutionary patterns in a large number of gene families, the logic extension of accomplished and ongoing projects are moved to:

- Develop more complex statistical models, for example, relaxation of some assumptions to be closer to the reality of biosystems (under way).
• Extend the study of functional divergence into the genetic architecture, for example, in terms of molecular pathways such as apoptotic pathways (preliminary results available).

• Explore the functional divergence of homologous genes by investigating expression profiles (results not shown in this dissertation).

• Infer the impact of gene duplication on the genome evolution by incorporating information such as chromosomal synteny, GC content (under way).
ACKNOWLEDGEMENTS

During the course of my Ph.D., there have been many people that have helped me to reach this milestone in my life. I would like to take this opportunity to recognize some of these people.

I wish to express my sincere appreciation to my primary advisor, Professor Xun Gu for his support and encouragement throughout the course of my research and studies. His guidance and insight were invaluable in completing my research and his patience during some of the stumbling blocks was greatly appreciated. Special acknowledgement is expressed to Professor Daniel Ashlock for his continuous support and advice. I would also like to thank other committee members Professors Linda Ambrosio, Kristen Johansen, and Gavin Naylor for their support during my Ph.D. I also want to thank my officemates, Ms. Jianying Gu and Ms. Molly Vanhouten. Our discussions of research, teaching, coursework, and life-in-general were always a welcome addition.

A deepest gratitude goes out to my parents. They have contributed so much to my life that I could never acknowledge it all in this limited space.
VITA

NAME OF AUTHOR: Yufeng Wang

DEGREES AWARDED:
   B.S. in Genetics, Fudan University, 1993
   M.S. in Statistics and Genetics (Co-major), Iowa State University, 1998

HONORS AND AWARDS:
   Graduate Teaching Excellence Award, Iowa State University, 1998
   James Cornette Research Excellence Award in Bioinformatics, Iowa State University, 2001

PROFESSIONAL PUBLICATIONS: