Functional divergence and genome evolution of vertebrate protein kinases (kinome)

Jianying Gu
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/rtd
Part of the Computational Biology Commons, Genetics Commons, and the Molecular Biology Commons

Recommended Citation

This Dissertation is brought to you for free and open access by Digital Repository @ Iowa State University. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Digital Repository @ Iowa State University. For more information, please contact digirep@iastate.edu.
Functional divergence and genome evolution of vertebrate protein kinases (kinome)

by

Jianying Gu

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

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Xun Gu, Co-major Professor
Dan Nettleton, Co-major Professor
Linda Ambrosio
Gavin Naylor
Daniel Voytas

Iowa State University
Ames, Iowa
2003

Copyright © Jianying Gu, 2003. All rights reserved.
Graduate College
Iowa State University

This is to certify that the doctoral dissertation of

Jianying Gu

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
# TABLE OF CONTENTS

## CHAPTER I: GENERAL INTRODUCTION
- Introduction 1
- Dissertation Organization 3
- References 5

## CHAPTER II: NATURAL HISTORY AND FUNCTIONAL DIVERGENCE IN PROTEIN TYROSINE KINASES
- Abstract 8
- Introduction 8
- Methods 9
- Results and Discussion 11
- Conclusions 23
- Acknowledgement 23
- References 23

## CHAPTER III: EVOLUTIONARY PERSPECTIVES FOR FUNCTIONAL DIVERGENCE OF JAK PROTEIN KINASE DOMAINS AND TISSUE-SPECIFIC GENES
- Abstract 27
- Introduction 28
- Methods 29
- Results 31
- Discussion 40
- Acknowledgements 43
- References 43

## CHAPTER IV: EVOLUTIONARY ANALYSIS OF FUNCTIONAL DIVERGENCE IN TGF-β SIGNALING PATHWAY
- Abstract 47
CHAPTER I: GENERAL INTRODUCTION

Introduction

Over the past two decades, advances in biology have increased exponentially, leading to an enormous wealth of genomic (the quantitative study of genes, regulatory and noncoding sequences), transcriptomic (RNA and gene expression), and proteomic (protein expression) information. Novel ways of using this information are emerging, with the hope that many unsolved problems in biology and medicine can be more fully explored. Evidently, an integrative approach that combines computational, quantitative, and experimental analysis brings new knowledge, novel methods, and innovative technologies to engender a better understanding of complex biological systems and processes. Many of these post-genomic attempts are centered at elucidating the function of gene products and functional interacting networks by unveiling the evolutionary forces that have shaped the network organization.

Gene/domain duplication, functional divergence, site-specific rate shift, and age distribution of vertebrate gene families

A central component in vertebrate cellular and developmental networks is gene families, in which a group of genes with common ancestry display sequence similarity but exert divergent functions (Lundin 1993; Hughes 1994; Spring 1997; Lynch and Conery 2000). It has long been thought that duplication of individual genes, chromosomal segments, or entire genomes serves as a primary resource for the function novelties in a vast number of gene families such as HOX (Ohno, 1970; Sidow 1996; Li et al. 2001). Domain duplication/shuffling has also been postulated to play additional role in increasing the functional complexity in multi-domain gene-family networks (Doolittle 1995; Henikoff et al. 1997). Yet still under debate, the functional divergence observed in gene families largely relies on the diverging fates of duplicated genes: (1) non-functionalization, in which one copy retains the original function and the other copy accumulates deleterious mutations and becomes a functionless pseudogene; (2) neo-functionalization, in which one of the duplicates retains its original function while the other is under relaxed selection constraint and may acquire novel function (Li, 1983); (3) Alternatively, sub-functionalization, in which each of
duplicates retains complementary part of the ancestral functions (Force et al. 1999; Lynch and Force 2000).

Provided the importance of gene/domain duplication in gene family innovation, it is intriguing to quantitatively trace the changes of evolutionary constraints after a duplication event. There is a general belief that not all, but only a few amino acid residues contribute to functional divergence after gene/domain duplication (Golding and Dean 1998). How to identify these residues from the background evolutionary noise is a challenge. Initially, sequence similarity score was used as an index for the level of functional divergence, with a caveat of incompetence to detect the true functional divergence, because most amino acid changes are neutral or nearly neutral which are not directly related to functionality (Eisen 1998). Instead of sequence similarity, pattern of amino acid replacement, including changes in the rate of replacement, appears to be a better indicator for changes of functional/selective constraint, which is likely to be inversely related to the evolutionary rate of amino acid replacement (Li and Graur 1991). Apparently, algorithms based on empirical rules may not be sufficient to capture the complex stochastic nature of evolution (Fetrow and Skolnick 1998; Montelione and Anderson 1999; Irving et al. 2001). Hence, a statistical framework that maps the association between inferred site-specific rate shifts and changes in function after gene/domain duplication is important (Gu 1999; Gaucher et al. 2001; Gu 2001). In addition to the statistical modeling of the functional divergence after gene/domain duplication, inferring the age-distribution of gene families that arose from duplication events will provide a deeper insight in the evolution of fine-tuned hierarchical cell network. The observation of up to four paralogs of many Drosophila genes in the human genome provides indirect evidence for the classical (two-round) hypothesis of vertebrate genome duplication (Ohno 1970). However, over the decades, the distinct modes of large-scale genome duplication vs. continuous small-scale gene duplication are hotly debated (Martin 2001; Friedman and Hughes 2001; Spring 1997; Wolfe 2001; McLysaght et al. 2002; Gu et al. 2002). Our recent study of vertebrate gene families (Gu et al. 2002) revealed that both large- and small-scale gene duplications have significant contribution to build the current hierarchy of the human proteome. Also recent segmental duplications have been shown to be involved in building up the extreme dynamism and genomic fluidity over very short periods of
evolutionary time (Eichler 2001). This series of studies have shed significant light on complicated mechanisms underlying the current diversity of vertebrate genomes.

**Kinome: key machinery in cell network**

Among thousands of gene families, kinase superfamilies that catalyze the reversible protein phosphorylation play a central role in regulating basic functions including DNA replication, cell cycle control, gene transcription, protein translation, and metabolism. Protein phosphorylation is also required for more advanced functions in higher eukaryotes including cell, organ, and limb differentiation, cell survival, synaptic transmission, cell-substratum and cell-cell communication, and to mediate complex interactions with the external environment. Moreover, aberrant protein phosphorylation is commonly related with cancer and other human diseases. The completion of human genome sequence provides us an opportunity to decipher the molecular nature of kinase-mediated signal transduction machinery (Venter et al. 2001). A complement of 518 kinases, defined as kinome (Manning et al. 2002), was identified in the human genome, constituting about 1.7% of all human genes. All of the known *bona fide* protein kinases share a conserved catalytic kinase domain of ~270 amino acids. Based on their characteristic kinase domain, eukaryotic Protein Kinases (ePKs) can be divided into two distinct families, protein tyrosine kinases (PTKs) and protein serine/threonine kinases (PSKs). Tyrosine protein kinase, one major animal-specific subclass, is likely monophyletic, derived from a precursor of protein serine-threonine kinase during the course of animal evolution, while the major types of PSKs had existed in the early stage of eukaryotes (Hanks et al. 1988; Hanks and Hunter, 1995; Hunter 1995). In addition to ePKs, a group of atypical Protein Kinases (aPKs) possess biochemical kinase activity, despite the lack of sequence similarity to the ePK domain. Interestingly, the structural similarity of aPKs to the authentic ePK domains may account for their enzymatic activity (Yamaguchi et al. 2001).

**Dissertation Organization**

The objective of this study is to explore the functional divergence and evolutionary pattern in the vertebrate kinome, using a combinatorial bioinformatic and evolutionary approach. This dissertation is composed of three sections: a general introduction, four
chapters of research reports, each of which is in journal manuscript format, and a general conclusion. The core of the five interrelated reports includes:

Chapter II summarizes a comprehensive phylogenetic analysis on major PTK gene families. We found that the expanding of each PTK family was likely caused by gene or genome duplication event(s) that occurred before the emergence of bony fishes but after the vertebrate-amphioxus split. Our further investigation on the evolutionary pattern revealed that site-specific shifted evolutionary rate (altered functional constraint) is a common pattern in PTK gene family evolution (Gu 1999).

Chapter III is a case study in Jak kinase family. The study was focused on the statistical modeling and detection of functional divergence among duplicate genes. Our result showed that shifted selective constraints (or shifted evolutionary rates) between homologous JH1 and JH2 domains are statistically significant. Moreover, we found that after the (first) gene duplication, site-specific rate shifts between member genes Jak2/Jak3 and Jak1/Tyk are significant, indicating a consequence of functional divergence among these member genes.

Chapter IV extends the study to kinase-mediate network: TGF-β signal transduction pathway which is triggered by receptor serine/threonine kinases at the cell surface. We have statistically tested the functional divergence in major pathway components, targeted the critical amino acid residues responsible for functional divergence. Interestingly, altered functional constraints appeared to be a common pattern in the evolution of TGF-β pathway. Moreover, we have found the correlation between structural divergence and functional divergence in the extracellular ligand-binding domain of Type II receptor kinases.

Chapter V extends the study to the impact of evolutionary forces that drive the vertebrate kinome organization. The important evolutionary events including gene duplication, domain duplication/shuffling, gene loss, and gene nonfunctionalization, which might have contribution to the existing kinome, have been identified. The age distribution of human kinome shows that the major kinase mediated signaling is generated by domain shuffling/duplications in the early evolution of eukaryotes, and subsequent large scale genome duplication is involved in the emergence of tissue-specific or developmental stage-specific kinase isoforms in the early stage of vertebrate. Our study of kinase pseudogenes
further demonstrated the fates of duplicate gene: gain new functionality or loss original function to turn into pseudogenes.

References:


CHAPTER II: NATURAL HISTORY AND FUNCTIONAL DIVERGENCE IN PROTEIN TYROSINE KINASES

A paper accepted by Gene

Jianying Gu, Xun Gu

Abstract

Cellular signaling is the core of many biological processes including growth, differentiation, adhesion, motility and apoptosis. The protein tyrosine kinase (PTK) supergene family is the key mediator in cellular signaling in metazoans, directly associated with a variety of human diseases. All PTKs contain a highly conserved catalytic kinase domain, in spite of variable multi-domain structures. Within each PTK gene family, members exhibit functional divergence in substrate-specificity or temporal/tissue-specific expression, although their primary function is conserved. After conducting phylogenetic analysis on major PTK gene families, we found that the expanding of each PTK family was likely caused by gene or genome duplication event(s) that occurred before the emergence of teleosts but after the vertebrate-amphioxus split. We further investigated the evolutionary pattern of functional divergence after gene duplication in those gene families. Our results show that site-specific shifted evolutionary rate (altered functional constraint) is a common pattern in PTK gene family evolution.

1. Introduction

Protein kinases catalyze the transfer of a phosphate group from its donor to an acceptor. They can be divided into two major groups: protein tyrosine kinases (PTKs) and serine/threonine kinases (PSKs) (Hubbard and Till, 2000). Since PTKs play important roles in regulating intracellular signaling pathways and are responsible for the development of many cancers, they have served as drug targets for many different disease therapies (Blume-
Jensen and Hunter, 2001; Sridhar et al., 2000). Therefore, there are general interests to explore how much functional-related information can be obtained from molecular evolutionary analysis.

PTKs can be further divided into receptor tyrosine kinases (RTKs) and non-receptor (cytosolic) tyrosine kinases. Typically, a cascade of signals is initiated from an RTK after ligand binding, which leads to receptor dimerization, kinase activation, and autophosphorylation of tyrosine residues. These phosphorylated tyrosines then serve as docking sites for recruiting downstream signaling molecules, including non-receptor tyrosine kinases that can trigger a variety of cell responses. The extensive cross talk between PTK-triggered pathways increases the complexity of the signaling process (see Schlessinger, 2000 for a review).

The human genome project has opened new opportunity in the study of PTK-mediated signaling. The first draft of complete human sequence predicted a total of 106 PTKs, 58 are receptor kinases and 48 are non-receptor kinases (Venter et al., 2001). Despite that different PTK subfamilies have different domain structures and functions, all PTKs share a conserved kinase domain that is responsible for their catalytic functions and structures. To unveil the intrinsic functional diversity among PTKs, natural history of gene family expansion, and the underlying evolutionary mechanism, we performed extensive phylogenetic-based analysis on 27 PTK families. Our study may provide some new insights for understanding the complexity of signal-transduction networks in the animal kingdom.

2. Materials and Methods

2.1 Data collection

The full-length sequences of vertebrate PTK gene families used in this study were obtained from the Hovergen database (http://phil.univ-lyon1.fr/), with reference from literatures (Robinson et al., 2000; Blume-Jensen and Hunter, 2001). Exhaustive PSI-BLAST search against Non-redundant Protein databases at NCBI was performed to find homologous sequences in invertebrates such as D. Melanogaster and C. elegans, which were used as outgroups. The complete alignment of kinase domain of PTKs was obtained from the Pfam (http://pfam.wustl.edu/). The chromosome synteny of individual PTKs in human genome
was obtained by the map viewer at NCBI (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search/). The final data set including the following PTK families:

(I) Receptor tyrosine kinases: Axl, discoidin domain receptor (DDR), epidermal growth factor receptor (EGFR), ephrin receptor (EphR), Fibroblast growth factor receptor (FGFR), hepatocyte growth factor receptor (HGFR), insulin receptor (InsR), Leukocyte tyrosine kinase (Ltk/Alk), muscle-specific kinase (Musk), platelet-derived growth factor receptor (PDGFR), Ret, receptor orphan (Ror), Ros, Ryk, Tie, tropomyosin-related kinase (Trk), and vascular endothelial growth factor receptor (VEGFR);

(II) Non-receptor tyrosine kinases: Abelson tyrosine kinase (Abl), acetate kinase (Ack), C-terminal src kinase (Csk), focal adhesion kinase (Fak), fps/fes related kinase (Fer/Fes), fyn-related kinase (Frk), Janus kinase (Jak), Src, Spleen tyrosine kinase (Syk), and Tec.

2.2 Multiple Alignment and phylogenetic analysis

The multiple alignment of each gene family was obtained by ClustalX program (http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html), followed by manual adjustment. Local alignments were reconciled with the kinase domain defined in the Pfam database by Hidden Markov Models (HMMs). Phylogenetic trees were inferred by the Neighbor-joining (NJ) method using MEGA2.0 (http://megasoftware.net/) with Poisson distance. The robustness of tree topology to the tree making method was examined by PAUP4.0 (http://paup.csit.fsu.edu/) and PHYLIP (http://evolution.genetics.washington.edu/phylip.html/). Bootstrap analysis was carried out to assess support for individual node.

2.3 Time estimation of gene duplication events

A linearized neighbor-joining tree (Takezaki et al., 1995) was used to convert the (average) Poisson distance of protein sequences to the molecular time scale by using software package MEGA2.0. In this study, the divergence time of primate-rodent (80 million year ago, mya), mammal-bird (310 mya), mammal-amphibian (350 mya), and tetrapod-teleost (430 mya) and vertebrate-Drosophila split (830 mya) were employed as calibrations (Kumar and Hedges, 1998; Gu, 1998; Wang and Gu, 2000).
2.4 Type I functional divergence analysis

Type I functional divergence refers to the evolutionary process that results in altered functional constraints (or site-specific evolutionary rate shift) between two duplicate genes, regardless of the underlying evolutionary mechanisms (Gu, 1999). A statistical framework modeling the functional divergence was used to estimate the coefficient of functional divergence (θ), an indicator of the level of Type I functional divergence of two homologous gene clusters (Gu, 1999; Wang and Gu, 2001; Gu et al., 2002a). Rejection of null hypothesis H₀: θ = 0 provides statistical evidence for shifts in inferred evolutionary rate (or altered functional constraints) in duplicates.

Moreover, the sites with critical contribution to the overall functional divergence can be predicted by the posterior analysis. Let \( Q(k) = P_k(S_1|X) \) be the posterior probability of a site k being \( S_1 \) (functional divergence related status) when the amino acid configuration (X) is observed. Since the alternative status \( S_0 \) (functional divergence unrelated status), with posterior probability \( P_k(S_0|X) = 1 - P_k(S_1|X) \), means no altered functional constraint, the predicted residues are only meaningful when \( Q(k) > 0.5 \), in which case the posterior odds ratio \( R(S_1/S_0) = P(S_1|X)/P(S_0|X) > 1 \). A more stringent cut-off may be \( Q(k) > 0.67 \), or \( R(S_1/S_0) > 2 \). The computational software DIVERGE for Type I functional divergence analysis and prediction is available at http://xgu1.zool.iastate.edu/.

3. Results and Discussions

3.1 The natural history of protein tyrosine kinase gene families

PTKs form a large but diverse superfamily with the characteristic catalytic domain conserved across different gene families. The classification of these gene families is based upon ligand specificity, biological function and primary structure. Within each gene family, the member genes are conserved in the sequence as well as the domain structures, but distinct in tissue or developmental-stage or ligand specificity.

Gene (domain) duplication plays an important role in eukaryote evolution by providing the primary source for the evolutionary novelty (Ohno, 1970). Recently we (Gu et al., 2002b) have demonstrated that both large-scale (genome-wide) and small-scale duplications have significant contributions on the presence of numerous multigene families.
The phylogenetic tree based on the alignment of kinase domain sequences suggested that the whole set of PTKs was generated from a series of domain duplications in the early stage of animal evolution, as marked by the “ancient component” in Gu et al. (2002b). These gene (domain) duplications have given rise to the current complex network of signal-transducting in animals (Hanks et al., 1995; Suga et al., 1997).

Furthermore, phylogenetic tree for each gene family in the PTK superfamily has revealed the impact of genome-wide duplication(s) in the early stage of vertebrates (Gu et al., 2002b). Moreover, we found no evidence for the contribution of recently gene duplication (around or after mammalian radiation) to the origin of tissue-specific PTKs. Because of space limitation, in the following we only discuss two typical cases.

**EphR gene family**

To date, 14 highly related ephrin receptors have been identified in vertebrates. They can be divided into two classes according to the ligand-binding preference: EphA receptors (EphA1-A8) recognize glycosyl-phosphatidyl-inositol (GPI)-anchored ephrin-A ligands (A1-A5), whereas EphB receptors (EphB1-B6) preferentially bind to ephrin-B ligands (B1-B3) that span the membrane via a transmembrane domain (Gale et al., 1996). One exception is EphA4, a receptor that can bind and respond to B as well as A-class ephrins.

The phylogenetic tree of ephrin receptors was inferred based on the whole length amino acid sequences, while its root was tentatively determined by the phylogeny of PTK superfamily. It shows that the A1/A2 ephrin receptor group first branched-off, probably after the split between vertebrates and Drosophila (Fig.1). The subsequent gene duplication has given rise to the two distinct classes of ephrin receptors; the rest of Class A and all the Class B ephrin receptors fall into separate clades with 93% bootstrap value. In each class more tissue-specific isoforms were generated by substantial gene duplications. Overall, the evolution and diversity of ephrin receptors was driven by both small-scale and large-scale gene duplications in the early stage of vertebrates. Of course, more vertebrate homologous genes (e.g., fishes) are needed to determine the phylogenetic interval of some duplication events. Nevertheless, molecular-clock analysis (Wang and Gu, 2000; Gu et al., 2002b) generally supports our notion (Fig.1).
Fig. 1. Phylogenetic tree of the Eph receptor gene family inferred by the neighbor joining method with poisson correction. Eph receptor family contains 14 highly related member genes which can be divided into two classes EphA receptors and EphB receptors with bootstrap value equal to 93%. Several duplication time points were indicated: (A) 403.1 mya; (B) 481.7 mya; (C) 526.0 mya; (D) 435.3 mya; (E) 477.7 mya; (F) 322.5 mya; (G) 596.6 mya; (H) 747.7 mya; (I) 784.0 mya.
If the rooting of EphR tree is largely correct, the evolution of ligand-binding preferences can be inferred from the phylogenetic analysis. That is, the ancestor of EphR may bind only ephrin-A ligands, while the function related to ephrin B ligands binding was derived relatively recently. This pattern is consistent with the classic theory of function innovation after gene duplication: one copy kept the original function (ephrin-A ligands) and the other one acquired novel function (ephrin-B ligands) through accumulation of amino acid change (Li, 1983). Interestingly, a similar pattern has also been observed within Class B ephrin receptors: Eph B2 and Eph B3 have been demonstrated to possess partial functional redundancy in midline guidance of nerve system, implying their common feature in preserving ancestral functions of Class B receptors (Cowan et al., 2000). Noticeably Eph B6 shows remarkably long branch length. One possibility is that it has undergone loss-of-function divergence due to the loss of crucial sites for tyrosine kinase activity and the relaxed functional constraints (Gurniak et al., 1996).

Src gene family

Src cytoplasmic tyrosine kinase family plays crucial roles in a variety of cellular processes, such as cell cycle control, cell adhesion, cell motility, cell proliferation and cell differentiation (Thomas et al., 1997). Extensive studies indicate that the complexity of functional roles of Src kinases mainly comes from their capability of communicating with a large number of upstream receptors and downstream effectors. The NJ tree shows that member genes of the Src family can be divided into two major distinct groups: (A) Src, Yes, Fyn, Fgr and Yrk; and (B) Lyn, Hck, Lck and Blk (Fig. 2). The three nearest neighbors to the Src family, tyrosine kinases Frk, Brk and Srm, were used to root the NJ tree.

There exists distinct differences between group A and B at gene expression level. The three group A member genes (Src/Fyn/Yes) are ubiquitously expressed; whereas all the group B member genes (Lck/Hck/Lyn/Blk) and one group A member gene (Fgr) have more tissue-restricted expression, mainly in hematopoietic cells. Research also showed significant structural difference between group A and B Src genes. The resolved crystal structures of human Src (Group A) and human Hck (Group B) reveal the significant geometrical difference between these two groups of proteins (Williams et al., 1998). Furthermore, the observation of swapping the SH2 and SH3 domains of Lck (group B) with the corresponding
Fig. 2. Phylogenetic structure of the Src gene family. The Src family can be divided into two major distinct groups: group A including Src, Yes, Fyn, Fgr and Yrk; group B including Lyn, Hck, Lck and Blk. Kinases Frk, Brk and Srm are used as outgroups to root the NJ tree. Identified chromosomal locations of member genes are listed according to human genome map.
domains of Src (group A) abolished the regulation of Src activities (Gonfloni et al., 1997) confirmed the non-replaceable functions. It is noteworthy that the N-terminus SH4 domain is very short and ubiquitously conserved for all Src member genes, indicating its common functional roles in the long run of evolution.

3.2 Evidence of gene duplication in PTK

The recent mounting evidence suggested that, in addition to continuous flux of small-scale gene duplications, there was at least one genome duplication event that occurred during early chordate evolution (Gu et al., 2002b; McLysaght et al., 2002). It is of particular interest to examine whether PTK gene families were generated through such gen(om)e duplication(s).

Age distribution of gene duplication in PTK families

Using the same approach as described by Gu et al. (2002b), we have dated the time of gene duplication events that gave rise to the PTK families. Among a total of 56 investigated duplication events, 38 occurred during the short time window 430–750 Mya which falls between the emergence of teleosts and the split of vertebrate-amphioxus (Fig. 3).

![The estimated time of gene duplications in PTK families](image)

Fig. 3. Age distribution of PTK gene duplication events. The molecular time scale is measured as Myr ago. Each bin for the histogram is 50 Myr.
Gu et al. (2002b) showed that both large- and small-scale duplications contributed to the current hierarchy of vertebrate genome: While a continuous mode of gene duplication is exhibited since the vertebrate-fruitfly split, a rapid increase in the number of paralogous genes was observed during the time period 430–750 Myr ago. The age distribution of PTK genes suggests the big-band explosion in the early vertebrate stage is a dominant force for shaping the diversity of functional specificity, where a continuous mode (constant rate of gene duplication) plays secondary role (chi-squared test with $p<0.0005$). Using a different approach, a similar conclusion was obtained by Miyata and his coworkers (Iwabe et al., 1996; see Miyata and Suga, 2001 for review).

Chromosome distribution of paralogous genes

Chromosome mapping of paralogous genes may provide useful information for the occurrence of genome duplication. We are aware that due to the frequent reshuffling and natural selection after chromosome duplications, it is difficult to reconstruct the real history of chromosome duplication events. Nevertheless, the present chromosome synteny still preserves some trace of the physical re-organization after gene duplication. For instance, a number of paralogous chromosomal regions (or paralogons) have been successfully identified recently (Popovici et al., 2001; McLysaght et al., 2002). Such paralogon in PTKs was reported as well: all the member genes of Fibroblast growth factor receptor (FGFR) family and seven other families were located in the same vicinity of chromosomal regions: 4p16, 5q33-35, 8p12-21, and 10q21-26 (Pebusque et al., 1998), indicating the possible occurrence of block duplication.

PDGFR/VEGFR gene family

The PDGFR/VEGFR family comprises by eight member genes: Pdgfr-$\alpha$ and Pdgfr-$\beta$, colony-stimulating factor 1 receptor (CSF1-R), stem cell factor receptor (SCFR, commonly known as c-Kit), Flt-3 (growth factor receptor for early hematopoietic progenitors), Vegfr-1, Vegfr-2, and Vegfr-3. This gene family seems to be vertebrate-specific since homologous sequence has not been found in invertebrates such as Drosophila and C. elegans. The analysis of phylogenetic hierarchy along with chromosome synteny provides striking evidence for the occurrence of the duplication of chromosome blocks during PDGFR family evolution. These eight member genes appear to have been arisen from a common ancestor,
based on their sequence similarity and their common structural features (e.g. high hydrophilic insertion into the catalytic kinase domain). Their diversification may be generated by a couple of gene duplication events. As shown in Figure 4, the first major duplication event took place in the early stage of vertebrates, at least prior to the divergence between tetrapods and teleosts. It involves the tetraploidization of three chromosome regions 13q12, 4q12, and 5q33-35. Consequently, one of the copies evolves to the present VEGFR isoforms 1, 2, and 3. Analogously, Flt3 may represent the other descendant in the region of 13q12. In contrast, complex changes have occurred in the close proximities of 4q12 (Kit and Pdgfr-α and counterparts) and 5q33-35 (CSF1R and Pdgfr-β) due to another round of gene duplication (Rousset et al., 1995). It is striking that the tandemly linked chromosome synteny is largely preserved over more than 400 million years. One plausible explanation is the crucial functional role of this group of receptor kinases decreases the possibility of radical chromosomal re-arrangement within the neighborhood.

**Src gene family**

In the Src gene family, the chromosome synteny also shows an interesting pattern. As shown in Figure 2, Src and Hck may be the descendant of a common ancestor located nearby human chromosome 20q, and Fgr and Lck may have arisen from the common ancestor in the proximity of chromosome 1p35-36. Though the entire puzzle of Src evolution remains unresolved, those observations lead us to hypothesize gene duplication (chromosome tetraploidization) as an important evolutionary mechanism underlying the organization of present-day Src genes.

### 3.3 Site-specific evolutionary rate shift in PTK gene families

We have conducted pair-wise functional divergence analysis between paralogous genes for each gene families. Gene clusters with less than four sequences were excluded from analysis due to insufficient information. Table 1 shows the coefficient of functional divergence (θ) of pair-wise comparisons of PTK superfamily. Forty-two (42) out of 43 comparisons showed θ > 0 with p < 0.05, suggesting that site-specific rate shift after gene duplication is a common phenomenon in PTKs evolution.

We have noticed an association between predicted functional divergence (via site-specific rate-shift) and observed (experimental or phenotypic) trait. For example, in the Src
Fig. 4. Phylogenetic tree of the PDGFR gene family. Phylogenetic hierarchy along with chromosome synteny provides evidence for the occurrence of the duplication of chromosome blocks during the PDGFR family evolution. The chromosome region 13q12 (Flt3 and VEGFR), 4q12 (Kit, PDGFR-α, and VEGFR2), and 5q33-35 (CSF1-R, PDGFR-β and VEGFR3) might be generated from a common ancestral chromosomal block following a series of chromosome duplications.
Table 1: The coefficient of functional divergence (θ) of pairwise comparisons of tissue-specific genes of protein tyrosine kinase gene families.

<table>
<thead>
<tr>
<th>Gene family name</th>
<th>Pairwise comparison²</th>
<th>θ ± se</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Receptor Kinase</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epidermal Growth Factor Receptor (EGFR)</td>
<td>Egfr (5) EgB2 (4)</td>
<td>0.306±0.084</td>
</tr>
<tr>
<td></td>
<td>Egfr (5) EgB3/4 (6)</td>
<td>0.184±0.046</td>
</tr>
<tr>
<td></td>
<td>EgB2 (4) EgB3/4 (6)</td>
<td>0.132±0.103</td>
</tr>
<tr>
<td>Insulin receptor (InsR/Igr-1R/Irr)</td>
<td>InsR (7) Igr-1R (8)</td>
<td>0.137±0.041</td>
</tr>
<tr>
<td>Platelet-derived growth factor receptor (PDGFR)</td>
<td>Kit (14) Fms (5)</td>
<td>0.215±0.032</td>
</tr>
<tr>
<td></td>
<td>Kit (14) Pdgfr (9)</td>
<td>0.161±0.026</td>
</tr>
<tr>
<td></td>
<td>Kit (14) Vegfr (11)</td>
<td>0.287±0.031</td>
</tr>
<tr>
<td></td>
<td>Fms (5) Pdgfr (9)</td>
<td>0.098±0.035</td>
</tr>
<tr>
<td></td>
<td>Fms (5) Vegfr (11)</td>
<td>0.342±0.043</td>
</tr>
<tr>
<td></td>
<td>Pdgfr (9) Vegfr (11)</td>
<td>0.233±0.032</td>
</tr>
<tr>
<td>Fibroblast growth factor receptors (FGFR)</td>
<td>Fgfr1 (7) Fgfr2 (9)</td>
<td>0.257±0.075</td>
</tr>
<tr>
<td></td>
<td>Fgfr1 (7) Fgfr3 (9)</td>
<td>0.407±0.066</td>
</tr>
<tr>
<td></td>
<td>Fgfr1 (7) Fgfr4 (8)</td>
<td>0.322±0.061</td>
</tr>
<tr>
<td></td>
<td>Fgfr2 (9) Fgfr3 (7)</td>
<td>0.210±0.074</td>
</tr>
<tr>
<td></td>
<td>Fgfr2 (9) Fgfr4 (8)</td>
<td>0.342±0.075</td>
</tr>
<tr>
<td></td>
<td>Fgfr3 (7) Fgfr4 (8)</td>
<td>0.338±0.060</td>
</tr>
<tr>
<td>Trk</td>
<td>TrkB (5) TrkC (5)</td>
<td>0.527±0.091</td>
</tr>
<tr>
<td></td>
<td>TrkB (5) Trk-related (4)</td>
<td>0.754±0.063</td>
</tr>
<tr>
<td></td>
<td>TrkC (5) Trk-related (4)</td>
<td>0.782±0.077</td>
</tr>
<tr>
<td>Hepatocyte growth factor receptor</td>
<td>Met (6) Ron (6)</td>
<td>0.222±0.037</td>
</tr>
<tr>
<td>Ephrin receptor</td>
<td>EphA (23) EphB (23)</td>
<td>0.159±0.031²</td>
</tr>
<tr>
<td>TIE/RET</td>
<td>Tie (8) Ret (6)</td>
<td>0.444±0.056</td>
</tr>
<tr>
<td>AXL</td>
<td>Mer (4) Tyr3 (5)</td>
<td>0.298±0.058</td>
</tr>
<tr>
<td></td>
<td>Mer (4) Ror1/2 (4)</td>
<td>0.763±0.092</td>
</tr>
<tr>
<td></td>
<td>Tyro3 (5) Ror1/2 (4)</td>
<td>0.796±0.094</td>
</tr>
<tr>
<td>SRC</td>
<td>Src (6) Yes (6)</td>
<td>0.283±0.095</td>
</tr>
<tr>
<td></td>
<td>Src (6) Fyn (8)</td>
<td>0.362±0.073</td>
</tr>
<tr>
<td></td>
<td>Src (6) Lck (4)</td>
<td>0.728±0.094</td>
</tr>
<tr>
<td></td>
<td>Src (6) Lyn (7)</td>
<td>0.728±0.098</td>
</tr>
<tr>
<td></td>
<td>Src (6) Brk/Srm (4)</td>
<td>0.466±0.163</td>
</tr>
<tr>
<td></td>
<td>Yes (6) Fyn (8)</td>
<td>0.424±0.098</td>
</tr>
<tr>
<td></td>
<td>Yes (6) Lck (4)</td>
<td>0.871±0.112</td>
</tr>
<tr>
<td></td>
<td>Yes (6) Lyn (7)</td>
<td>0.661±0.111</td>
</tr>
<tr>
<td></td>
<td>Yes (6) Brk/Srm (4)</td>
<td>0.431±0.222</td>
</tr>
<tr>
<td></td>
<td>Fyn (8) Lck (4)</td>
<td>0.642±0.081</td>
</tr>
<tr>
<td></td>
<td>Fyn (8) Lyn (7)</td>
<td>0.768±0.094</td>
</tr>
<tr>
<td></td>
<td>Fyn (8) Brk/Srm (4)</td>
<td>0.566±0.088</td>
</tr>
<tr>
<td></td>
<td>Lck (4) Lyn (7)</td>
<td>0.582±0.102</td>
</tr>
<tr>
<td></td>
<td>Lck (4) Brk/Srm (4)</td>
<td>0.315±0.127</td>
</tr>
<tr>
<td></td>
<td>Lyn (7) Brk/Srm (4)</td>
<td>0.246±0.179</td>
</tr>
<tr>
<td><strong>Non-receptor kinase</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Janus kinase</td>
<td>Jak1 (7) Jak2 (6)</td>
<td>0.351±0.041</td>
</tr>
<tr>
<td></td>
<td>Jak1 (7) Jak3 (6)</td>
<td>0.280±0.040</td>
</tr>
<tr>
<td></td>
<td>Jak2 (6) Jak3 (6)</td>
<td>0.361±0.034</td>
</tr>
</tbody>
</table>

a) The number inside the parentheses represents the number of sequences for that member gene.

family, $\theta$ between two major groups A and B is estimated to be $0.290 \pm 0.045$, suggesting substantial alterations in their site-specific selective constraints, reflected by the difference of expression, structure and functional role between two groups. In particular, in group A genes, $\theta$ between Fyn and Src is estimated to be $0.362 \pm 0.073$. This functional divergence is well demonstrated by the knock-out assay: $\textit{fyn}^{-}$ mice have alterations in a specific region of brain, the hippocampus, where some layers have more neurons than wild type mice and cause the respective layers to undulate; in contrast, $\textit{src}^{-}$ mice resembled wild type controls (Grant et al., 1992).

Amino acid residues responsible for such functional divergence after gene duplication can be predicted based on a site-specific profile, which represents the posterior probability of a site being functional divergence related status given the amino acid configuration. By choosing a suitable cut-off value, we will obtain a list of candidate sites, which could be mapped onto the secondary or domain structures. For example, in the Trk family, the posterior analysis predicts 27 amino acid sites critical for functional divergence between TrkB and TrkC, given the cutoff value $Q(k) > 0.7$ (Note: $\theta$ between TrkB and TrkC is estimated to be $0.527 \pm 0.091$). Three of predicted sites are located in the Leucine-rich domain, two in the first Ig-like domain, and two in the kinase domain. Interestingly, another ten sites reside inside the second Immunoglobulin (Ig)-like domain, which has been shown the dominant element for neurotrophin-binding specificity in the crystallographic studies (Urfer et al., 1998; Wiesmann et al., 1999). The spatial distribution of these predicted sites indicates that they are not equally dispersed through the whole coding region. Instead, the majority (17/27) are located inside the functional domains, indicating the importance of domain structure in the functional innovation after gene duplication.

The predicted sites also provide hint for the association between site-specific rate shift and molecular phenotypes, for example, as shown in hepatocyte growth factor receptor (HGFR) family. The HGFR family induces mitogenic, motogenic and morphogenic cellular response, as well as tumorigenesis by triggering a multi-step genetic program called 'invasive growth' including cell-dissociation, invasion of extracellular matrices and growth. Two member genes (Met and Ron/Sea) have been identified through human to Fugu, a teleost fish (Fig. 5a). The functional divergence analysis shows significant changes in the functional
Fig. 5. Functional divergence analysis of HGFR gene family. The overall coefficient of functional divergence between Met and Ron is $\theta \pm se = 0.222 \pm 0.037$. (a) Phylogenetic tree of HGFR gene family. (b) The site-specific profile of posterior probability of functional divergence between Met and Ron. (c) Amino acid configuration of sites with $Q(k) > 0.6$. Predicted sites can be divided into two categories: category I contains sites conserved in Ron but variable in Met; and category II includes sites conserved in Met but variable in Ron.
constraints between Met and Ron ($\theta = 0.222 \pm 0.037$). Furthermore, posterior analysis implies strong variation in evolutionary rate shift of each site. As shown in figure 5b, the baseline of posterior probability profile is between 0.2–0.3, suggesting that most amino acid residues do not have significant effects on the overall functional divergence. Only a small portion of residues has high posterior probability to be responsible for the functional divergence between Met and Ron. With the cut-off value $Q(k) > 0.6$, 11 sites were predicted, which can be divided into two categories: category I contains sites conserved in Ron but variable in Met; and vice versa, category II includes sites conserved in Met but variable in Ron (Fig. 5c). Given the observation that the knock-out mice lacking *met* have placental-lethal defect, but *ron* knockout embryos are viable through the blastcyst stage of development (Uehara et al., 1995; Muraoka et al., 1999), we speculate that Met has indispensable functions and is likely under more stringent functional constraint than Ron. Indeed, the highly conservation of these category II sites in Met may be very important to maintain the specific Met function.

4. Conclusions

In conclusion, our comprehensive evolutionary analysis on PTKs reveals that (1) both large-scale and small-scale gene duplications are the major evolutionary force for generating the contemporary PTK superfamily. (2) Substantial functional divergence occurred after gene duplication(s), characterized by the significant shift in evolutionary rates. (3) Evolutionary functional divergence is correlated with the phenotypic functional divergence in paralogous genes. These results not only shed light on the role of gene duplication in the development of hierarchical PTK-mediated network, but also provide impetus for a new approach to predict functionary divergence from evolutionary changes.

Acknowledgement

We thank Yufeng Wang for helpful discussion. This study is supported by the NIH grant RO1 GM62118 to X. G.

References


Abstract

Jak (Janus kinase) is a non-receptor tyrosine kinase, which plays important roles in signal transduction pathways. The unique feature of Jak is that, in addition to a fully functional tyrosine kinase domain (JH1), Jak possesses a pseudokinase domain (JH2). Although JH2 lost its catalytic function, experimental evidence has shown that this domain may have acquired some new but unknown functions. This apparent functional divergence after the (internal) domain duplication may result in dramatic changes of selective constraints at some sites. We have conducted a data analysis to test this hypothesis. Our result shows that shifted selective constraints (or shifted evolutionary rates) between JH1 and JH2 domains are statistically significant. Predicted amino acid sites by posterior analysis can be classified into two groups: very conserved in JH1 but highly variable in JH2, and vice versa. Moreover, we have studied the evolutionary pattern of four tissue-specific genes, Jak1, Jak2, Jak3 and Tyk2, which were generated in the early stages of vertebrates. We found that after the (first) gene duplication, site-specific rate shifts between Jak2/Jak3 and Jak1/Tyk are significant, presumably as a consequence of functional divergence among these genes. The implication of our study for functional genomics is discussed.

Introduction

1 Reprinted with permission of Journal of Molecular Evolution, 2002, 54, 725-733.
2 Graduate student and Associate Professor, respectively, Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, Program of Bioinformatics and Computational Biology (BCB), Iowa State University.
3 Primary researcher and author
4 Author for correspondence
Jak protein, a group of non-receptor tyrosine kinases, plays a crucial role in the cytokine signaling (Darnell et al. 1994; Leonard and O’Shea 1998). The Jak family consists of four mammalian members: Jak1, Jak2, Jak3, and Tyk2 (Wilks et al. 1991; Harpur et al. 1992; Rane and Reddy 1994). Homologous Jaks from aves and teleosts, as well as a Drosophila Jak homologue, hopscotch, have also been sequenced (Binari and Perrimon 1994, Conway et al. 1997). The distinguishing feature of the Jak protein family is the existence of tandem kinase (JH1) and pseudokinase (JH2) domains. The tandem kinase domain has all the functional features of a typical tyrosine kinase domain, while the pseudokinase domain has lost the catalytic activity (Guschin et al. 1995; Feng et al. 1997; Zhou et al. 1997).

Among metazoan protein tyrosine kinases (PTKs), only Jak protein maintains a pseudokinase domain (JH2). The reason for this phenomenon remains unclear. JH2 has all the subdomains that are shared by typical tyrosine kinase domains, indicating a common origin with other kinase domains, e.g., via internal domain duplication. Thus, these two domains in Jak proteins provide an interesting case study of functional divergence in proteins from an evolutionary perspective. It is believed that the lack of some typical motifs, shared by functional kinase domains, in the pseudokinase domain rendered this domain catalytically inactive (Wilks et al. 1991; Frank et al. 1995). On the other hand, knock-out evidence has shown either abrogation or stimulation of kinase activity when the JH2 domain is deleted in Jak2, Jak3 or Tyk2 (Velazquez et al. 1995; Candotti et al. 1997; Luo et al. 1997). These facts imply that after the internal domain duplication, the pseudokinase domain has undergone loss-of-function and gain-of-function simultaneously. An interesting question in molecular evolution is whether functional alterations in the JH2 domain result in considerable changes in selective constraints (i.e., different evolutionary rate) at those sites involved. As a consequence, some well-conserved sites in JH1 domain become highly variable in the JH2 domain, and vice versa. Such a pattern has been described as “covarion behavior” (Gaucher et. al 2001), or Type I functional divergence (Gu 1999).

In this paper, we use Gu’s (1999) method to test the hypothesis that altered selective constraints at some positions occurred between the JH1/JH2 domains. Then, we studied the evolutionary pattern among vertebrate Jak tissue-specific genes, and compared this to the JH1/JH2 domain evolution. The implication of our results for functional divergence and specification in Jak/Stat pathway is discussed.
Methods

Datasets

We searched the Pfam database (http://pfam.wustl.edu/) for sequences that have tandem kinase or pseudokinase domains of Jak genes. For comparison, the kinase domain sequences of FGFR (Fibroblast Growth Factor Receptor) and EGFR (Epidermal Growth Factor Receptor) gene families were also downloaded.

To find all available sequences that belong to Jak gene family, the Gapped BLAST and PSI-BLAST searches were performed in several protein databases by using the human Jak1 gene as a query sequence (Altschul et al. 1997). After the exhaustive search, partial sequences and redundant sequences were removed from further analysis. The final data set includes 22 complete vertebrate Jak sequences and 1 Drosophila homologous gene Hopscotch (see figure 4 legend for accession number and species).

Multiple Alignment and Phylogenetic Analysis

The multiple hidden Markov Models (HMM) alignment of tandem kinase and pseudokinase domains was obtained from Pfam, followed by manual editing. The multiple alignment of complete Jak sequences was obtained using the program Clustal X (Thompson et al. 1997). These multiple alignments are available upon request. A phylogenetic tree was inferred by the neighbor-joining method (Saitou and Nei 1987) using the software MEGA2.0 (http://www.megasoftware.net/) (Kumar et al. 1994). Parsimony (PAUP) and likelihood methods (PHYLIP) were also used for phylogenetic reconstruction to examine sensitivity to tree-making method.

The ratio of nonsynonymous rate to synonymous rate (dn/ds) was calculated by using a modified version of Gojobori and Nei's method (in the software MEGA2.0). When assuming that the synonymous substitution is virtually neutral, dn/ds > 1 indicates positive selection, dn/ds < 1 indicates negative selection, and dn/ds = 1 indicates neutral evolution.

Testing Type 1 functional divergence (altered selective constraint)

Type 1 functional divergence refers to the evolutionary process that results in altered selective constraints (different evolutionary rates) between two duplicate genes, regardless of
the underlying evolutionary mechanisms (Gu 1999, 2001). Its functional-structural basis was well illustrated in the case study of Caspase family (Wang and Gu 2001). Consider two gene clusters generated by duplication. Briefly speaking, in each cluster an amino acid site is called an $F_1$-site (functional divergence-related) if its evolutionary rate differs from that in the ancestral gene; otherwise it is called an $F_0$-site (i.e., functional divergence-unrelated). Consequently, in the case of two gene clusters, a site can be in either of two states: (1) $S_0$, or (2) $S_1$, where $S_0$ is defined as a site being $F_0$ in both clusters and $S_1$ is defined as a site being $F_1$ in at least one cluster. The coefficient of Type I functional divergence, $	heta_{A,B}$, between two gene clusters A and B is defined as the probability of a site being in the $S_1$ state. Thus, a null hypothesis of $\theta = 0$ means that the evolutionary rate is virtually the same between duplicate genes at each site. Note that many models for rate variation among sites are the special case of $\theta = 0$. A probabilistic model has been developed (Gu 1999) to estimate $\theta$ by establishing the relationship between this type of functional divergence and the observed amino acid configurations.

Rejection of the null hypothesis (i.e., $\theta > 0$) provides statistical evidence for altered selective constraints of amino acid sites after gene duplication. Thus, we can use a site-specific profile to identify responsible amino acid sites. Specifically we define $Q_k$ to be the posterior probability that site $k$ is in state $S_1$ ($0 \leq Q_k \leq 1$). Large $Q_k$ indicates a high possibility that the functional constraint (or, the evolutionary rate) of a site is different between two clusters.

The software DIVERGE (http://xgu1.zool.iastate.edu) was used for the functional divergence analysis.

**Functional distance analysis**

One limitation of the two-cluster analysis described above is that it cannot test whether one duplicate gene has more shifted evolutionary rates than the other one. This issue is addressed by a simple method described below that can be applied when more than two homologous gene clusters are available (Wang and Gu 2001). First, the Type I functional distance between any two clusters is defined as $d_F = -\ln (1-\theta)$. Under the assumption of independence, Wang and Gu (2001) have shown that $d_F$ is additive, i.e., for two clusters A and B, $d_F(A,B) = b_F(A) + b_F(B)$, where $b_F(x)$ is the functional branch length of a given gene.
cluster x. Large $b_F$ value for a gene cluster indicates the evolutionary conservation may be shifted at many sites.

The estimated coefficients of Type I functional divergence ($\theta$) for all the pairs of clusters can be used to create a matrix of $d_F$ values. Given this matrix, a standard least squares method can be implemented based on the formula $d_F(A,B) = b_F(A) + b_F(B)$ to estimate $b_F$ for each gene cluster. If $b_F = 0$, it indicates that the evolutionary rate of each site in this duplicate gene has remained nearly the same since the gene duplication event, indicative that the derived state is more similar to the ancestral state for this particular cluster.

Results

Pattern of functional divergence between tandem kinase and pseudokinase domains

*The kinase and pseudokinase domains in Jaks*

In addition to the regular kinase domain (JH1), Jak proteins contain a pseudo-kinase domain (JH2), with a functional role that has not been clearly determined (Aringer et al. 1999). To explore the evolutionary pattern of JH1 and JH2 domains, we reconstructed a neighbor-joining (NJ) tree, including Jaks and two closely related protein tyrosine kinases, FGFR and EGFR. The inferred phylogeny shows that tandem kinase (JH1) and pseudokinase (JH2) domains are evolutionarily distinct (Figure 1). Indeed, tandem kinase domains (JH1) in Jaks appear to be more closely related to the functional kinase domains of FGFRs and EGFRs, while the pseudokinase domains (JH2) of Jaks form a unique clade. In fact, we have found that the JH2 domain had been generated before the emergence of most member genes of the Protein tyrosine kinase supergene family (data not shown).

It has been suggested that the pseudokinase domains (JH2) no long exhibit the catalytic activity, but may have acquired some new functions (Aringer et al. 1999). Thus, it is interesting to test whether this functional divergence resulted in shifted selective constraints (different evolutionary rates) at some sites between tandem kinase (JH1) and pseudokinase (JH2) domains. To this end, we estimated the coefficient of functional divergence ($\theta$) between JH1, JH2, EGFR and FGFR domains, as shown in the upper diagonal of Table 1. All of the estimates are significantly greater than 0 ($p < 0.01$). In particular, $\theta_{JH1, JH2} \pm se = 0.412 \pm 0.049$ provides statistical evidence for supporting the hypothesis of altered selective constraints between tandem kinase (JH1) and pseudokinase (JH2) domains in Jak proteins.
Figure 1. 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).
Table 1: $\theta_{AB}$ 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 $\theta_{AB}$ for all the pair-wise comparisons for Jaks, FGFRs, and EGFRs, where $\theta_{AB}$ is defined as the coefficient of the functional divergence between cluster $A$ and cluster $B$. The lower diagonal shows $d_{f}(A,B)=-\ln(1-\theta)$ which is defined as the functional distance between cluster $A$ and cluster $B$.

To further test the hypothesis that JH2 is more functionally divergent than JH1, we conducted the functional distance analysis (Wang and Gu 2001). The functional distances between domains [$d_{f} = -\ln (1-\theta)$] are shown in the lower diagonal of Table 1. Subsequently, the functional branch length ($b_{f}$) for each domain was estimated by the least-squares method, which can be illustrated as a star-like tree (Figure 2). The null hypothesis of equal functional branch lengths is rejected ($p < 0.05$). Interestingly, the functional branch length of the pseudokinase domain (JH2) of Jak proteins is about four times greater than that of the kinase domain (JH1), indicating that the distinct functional role of JH2 is likely generated by the episodic evolution of kinase domains.

Important amino acid sites for 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 sites that are responsible for functional divergence between tandem kinase (JH1) and pseudokinase (JH2) domains. Among 212 amino acid sites, 154 (73%) sites have $Q_k < 0.5$ implying low probability of contribution to functional divergence. For the remaining 58 amino acid sites with $Q_k > 0.5$, 21 of them show a very high probability of being functional divergence-related [$Q_k > 0.9$]. These 21 sites can be definitively grouped into two
Figure 2. The tree-like topology of kinase domains of Jaks, FGFRs, and EGFRs in terms of the functional distance $b_F$.

Table 2. Predicting critical amino acid sites responsible for functional divergence between tandem kinase and pseudokinase domain in Jaks.

<table>
<thead>
<tr>
<th>I Position(k)</th>
<th>$Q_k$</th>
<th>II Position(k)</th>
<th>$Q_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>0.997</td>
<td>38</td>
<td>0.998</td>
</tr>
<tr>
<td>203</td>
<td>0.992</td>
<td>47</td>
<td>0.967</td>
</tr>
<tr>
<td>30</td>
<td>0.975</td>
<td>41</td>
<td>0.960</td>
</tr>
<tr>
<td>20</td>
<td>0.968</td>
<td>190</td>
<td>0.959</td>
</tr>
<tr>
<td>46</td>
<td>0.967</td>
<td>105</td>
<td>0.956</td>
</tr>
<tr>
<td>23</td>
<td>0.966</td>
<td>103</td>
<td>0.954</td>
</tr>
<tr>
<td>162</td>
<td>0.963</td>
<td>44</td>
<td>0.931</td>
</tr>
<tr>
<td>70</td>
<td>0.959</td>
<td>121</td>
<td>0.918</td>
</tr>
<tr>
<td><strong>137</strong></td>
<td><strong>0.957</strong></td>
<td>87</td>
<td>0.906</td>
</tr>
<tr>
<td>21</td>
<td>0.922</td>
<td></td>
<td></td>
</tr>
<tr>
<td>159</td>
<td>0.913</td>
<td></td>
<td></td>
</tr>
<tr>
<td>196</td>
<td>0.906</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Category I: conserved in the tandem kinase domain; variable in the pseudokinase domain; Category II: conserved in the pseudokinase domain; variable in the tandem kinase domain.
categories: (I) conserved in the tandem kinase (JH1) domain, but variable in the pseudokinase (JH2) domain; (II) conserved in the pseudokinase domain, but variable in the tandem kinase domain (Table 2).

Category I: Of the 12 sites belonging to this category, one site, site 137 ($Q_{137} = 0.957$), has been demonstrated to be a determining site for the function of the tandem kinase domain (JH1), 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 site 137 is invariant in tandem kinase (JH1) domains. In contrast, the same position in pseudokinase domains (JH2) has a variety of amino acids with very different chemical properties. For example, some JH2 domains have amino acids with non-polar side chains such as glycine, alanine and proline, and some of them have uncharged polar amino acids like serine and threonine (Figure 3A). This observation can be explained as a relaxed selective constraint that was caused by loss-of-function in phosphorylation in JH2 domains.

Category II: Nine predicted sites belong to this category (Figure 3B). Among them, site 103 is predicted to be highly functional-divergence related [$Q_{103} = 0.954$]. Experimental data show that a glutamic acid (E)-to-lysine (K) replacement occurring at this site in the pseudokinase (JH2) domain hyperactivated Jak-Stat pathway in Drosophila and mammalian species (Luo et al. 1997). It seems likely that after the internal domain duplication, the tandem kinase domain (JH1) largely maintained the original catalytic function, while the pseudokinase domain (JH2) may have achieved some unidentified new functions, resulting in a set of JH2-specific conserved sites.

**Pattern of functional divergence between Jak member genes**

*Functional divergence of Jak gene family*

Figure 4 shows the neighbor-joining (NJ) tree of four Jak gene members (Jak1, Jak2, Jak3 and Tyk2). Parsimony method (PAUP) and likelihood method (PHYLIP) give essentially the same topology (data not shown). The phylogenetic analysis indicates that Jak member genes were generated by two gene duplications in the early stage of vertebrates, i.e.,
**Figure 3.** Functional divergence related amino acid sites 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 phylogenetic tree of the Jak gene family. The neighbor-joining algorithm was used to infer the topology based on the multiple sequence alignment with Poisson distance. Bootstrap scores >50% are presented. 1st and 2nd represent the time points of two rounds of gene duplications. Accession numbers: **Jak2**: L16956(mouse); U13396(rat); AF005216(human); AB006011(pig); AF090382(pufferfish); AJ005690(zebrafish). **Jak3**: U09607(human); L40172(mouse); D28508(rat); AF034576(chicken); AF148993(carp); AF091238(pufferfish). **Tyk2**: X54637(human); AF173032(mouse); AF090383 (pufferfish). **Jak1**: M64174(human); AB036335(pig); M33425(mouse); AF096264 (chicken); L24895(carp); U82980(zebrafish); U53213(pufferfish). Hopscotch: L26975(*Drosophila)*
before the emergence of teleosts. The first gene duplication resulted in the common ancestor of Jak2/3 and Tyk2/Jak1, followed by the second one resulting in the current four member genes. Note that Wang and Gu (2000) have shown that many tissue-specific gene families show a similar pattern, raising a possibility of large-scale duplication(s) in early vertebrates.

As nucleotide change in the regulatory region after gene duplication may be the first step for duplicate gene preservation (Force et al. 1999), amino acid replacements are responsible for the functional divergence for specification among tissue-specific isoforms. To explore whether (site-specific) altered selective constraints occurred during Jak family evolution, we estimated $\theta$ between Jak member genes, as well as between two groups (Jak2/3 and Tyk2/Jak1) that was generated by the first round gene duplication. The estimation is based on the phylogenetic tree in figure 4. We have found that (1) (site-specific) altered selective constraint after the first gene duplication is statistically significant, and (2) the $\theta$ value between Jak2 and Jak3 is $0.019 \pm 0.059$, indicating overall no significant site-specific shift of evolutionary rate between them (Table 3A).

Furthermore, we have estimated the coefficients of Type I functional divergence between Jak genes for three separate regions: (1) the tandem kinase domain (JHI), (2) the pseudokinase domain (JH2), and (3) the surrounding region excluding the JHI and JH2 domains (Table 3B). Our results can be summarized as follows. First, after first-round gene duplication, site-specific altered selective constraint in the JHI domain is statistically significant ($p < 0.01$), but not significant in the JH2 domain; there is weak evidence in the surrounding region ($p \sim 0.05$). Second, after the second gene duplication, no statistical evidence is observed for altered selective constraint in any domain. Since the estimate of $\theta$ in the JHI domain tends to be higher than those of JH2 domain or the surrounding region, we conclude that JHI domain is mainly responsible for functional divergence among Jak tissue-specific genes.

**Isoform-specific functional divergence of Jak gene family**

The functional branch lengths ($b_F$) of Jak1, Jak2, and Jak3, can be estimated from the pair-wise estimates of $\theta$ between Jak1, Jak2, and Jak3, respectively. Table 4 shows the results for the whole length of Jak proteins, and three domains (JHI, JH2, and the surrounding region), respectively. Interestingly, the level of altered selective constraints of
Table 3. The pair-wise coefficient of functional divergence ($\theta_{AB}$) for different member genes of Jak gene family. (A) $\theta$ values for the full length Jak proteins; (B) $\theta$ values for the different regions of Jak proteins.

(A)

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Gene duplication</th>
<th>$\theta \pm \text{se}^{(2)}$ (ML)</th>
<th>LRT$^{(3)}$</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jak1</td>
<td>Jak2</td>
<td>I</td>
<td>$0.208 \pm 0.056$</td>
<td>13.609</td>
<td>&lt;0.0005</td>
</tr>
<tr>
<td>Jak1</td>
<td>Jak3</td>
<td>I</td>
<td>$0.158 \pm 0.051$</td>
<td>9.502</td>
<td>0.002</td>
</tr>
<tr>
<td>Jak2</td>
<td>Jak3</td>
<td>II</td>
<td>$0.019 \pm 0.059$</td>
<td>0.110</td>
<td>0.75</td>
</tr>
<tr>
<td>Jak1/Tyk2</td>
<td>Jak2/Jak3</td>
<td>I</td>
<td>$0.072 \pm 0.027$</td>
<td>7.292</td>
<td>0.007</td>
</tr>
<tr>
<td>Jak1</td>
<td>Jak2/Jak3</td>
<td>I</td>
<td>$0.147 \pm 0.039$</td>
<td>14.367</td>
<td>&lt;0.0005</td>
</tr>
</tbody>
</table>

Note: As a single gene cluster, Tyk2 was excluded from the analysis due to the insufficient number of sequences.

(B)

<table>
<thead>
<tr>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Gene duplication</th>
<th>JH1</th>
<th>JH2</th>
<th>Surrounding</th>
<th>Full length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jak1</td>
<td>Jak2</td>
<td>I</td>
<td>0.435±0.140</td>
<td>0.214±0.125</td>
<td>0.294±0.100</td>
<td>0.208±0.056</td>
</tr>
<tr>
<td>Jak1</td>
<td>Jak3</td>
<td>I</td>
<td>0.231±0.086</td>
<td>0.052±0.143</td>
<td>0.178±0.089</td>
<td>0.158±0.051</td>
</tr>
<tr>
<td>Jak2</td>
<td>Jak3</td>
<td>II</td>
<td>0.217±0.150</td>
<td>0.001±0.022</td>
<td>0.022±0.100</td>
<td>0.019±0.059</td>
</tr>
<tr>
<td>Jak1/Tyk2</td>
<td>Jak2/Jak3</td>
<td>I</td>
<td>0.122±0.050</td>
<td>0.094±0.068</td>
<td>0.101±0.053</td>
<td>0.072±0.027</td>
</tr>
<tr>
<td>Jak1</td>
<td>Jak2/Jak3</td>
<td>I</td>
<td>0.174±0.067</td>
<td>0.086±0.097</td>
<td>0.225±0.074</td>
<td>0.147±0.039</td>
</tr>
</tbody>
</table>

(1) I represents the first gene duplication; II stands for second gene duplication.
(2) se represent standard error.
(3) likelihood ratio statistic.

Table 4. Functional branch lengths ($b_F$) of different regions in three Jak isoforms.

<table>
<thead>
<tr>
<th></th>
<th>JH1</th>
<th>JH2</th>
<th>Surrounding</th>
<th>Full length</th>
<th>$d_s/d_a$ (Full length)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jak1</td>
<td>0.295</td>
<td>0.147</td>
<td>0.261</td>
<td>0.193</td>
<td>0.074</td>
</tr>
<tr>
<td>Jak2</td>
<td>0.277</td>
<td>0.095</td>
<td>0.087</td>
<td>0.040</td>
<td>0.089</td>
</tr>
<tr>
<td>Jak3</td>
<td>-0.032</td>
<td>-0.094</td>
<td>-0.065</td>
<td>-0.021</td>
<td>0.274</td>
</tr>
</tbody>
</table>
member genes, measured by $b_F$, follows $b_F(Jak1) > b_F(Jak2) > b_F(Jak3)$, while the level of altered selective constraints of domains follows JH1 > surrounding region > JH2. In particular, the $b_F$ for Jak3 is virtually zero for all three domains, whereas JH2 (the pseudokinase domain) shows no significant functional branch length in Jak2 and Jak3.

Table 4 also includes the ratios of nonsynonymous to synonymous rates ($dn/ds$) of Jak member genes, based on human-mouse orthologous members of a gene, which can be used to measure the difference of selective constraints among member genes of a gene family (Tsunoyama et al. 1998). Interestingly, Jak3, which has virtually $b_F = 0$, shows the highest $dn/ds$ ratio. This observed negative association between $dn/ds$ and $b_F$ can be interpreted as follows. In the early stage after gene duplication, the functional divergence may occur in one of two lineages (measured by large $b_F$ value). If this process leads to the acquisition of some new functions, a stronger functional constraint (measured by low $dn/ds$ value) is expected.

In summary, during Jak tissue-specific gene evolution, Type I functional divergence (i.e., site-specific altered selective constraints) occurs mainly in the JH1 domain of Jak1 and Jak2 genes. As Jak1 and Jak2 genes may have evolved specialized functions in their kinase domains, Jak3 is more likely to have inherited the ancestral function.

Predicting important amino acid sites for Type I functional divergence

Critical amino acid sites responsible for the Type I functional divergence among Jak member genes can be predicted by $Q_k$, the posterior probability of being functional divergence-related at site $k$ (Gu 1999). Figure 5 shows the site-specific profile between Jak1/Tyk2 and Jak2/Jak3 clusters, measured by $Q_k$. It clearly shows that, after the first gene duplication, only a small portion of sites have undergone shifted rates, as indicated by high posterior probabilities. These sites may be considered as prime candidates for further functional assays.

Discussion

Domain and gene duplications are both important for Jak gene family proliferation and evolution. After gene (or domain) duplication, one copy mainly retains the original function, whereas the other copy is under relaxed evolutionary constraint that may result in functional divergence and/or specification (Ohno 1970; Li 1983). In this paper, we have
investigated the evolutionary pattern of tandem kinase (JH1) and pseudokinase domains (JH2), a unique feature in the super-gene family of tyrosine protein kinases. Our results are summarized as follows.

In the very early stage of animals, an internal duplication had occurred within the ancestor of Jak gene, producing two kinase domains. As one domain (JH1) largely retains the original kinase activity, the other one (JH2) is free to accumulate amino acid replacements because of functional redundancy. When some key sites were replaced by other amino acids without serious deleterious effects on the survival, the kinase function of the JH2 domain was lost. Consequently, some well-conserved sites in functional kinase domains turn out to be highly variable. Meanwhile, the JH2 domain appears to have acquired some new functions that may explain its long-term existence during evolution. Despite the lack of substantial evidence for a physiological role of JH2, we indeed observed a group of sites in the JH2 domain to be very conserved, whereas they were variable in the JH1 domain.

It seems likely that the functional divergence between JH1 and JH2 domains had completed before the origin of vertebrates. In the early vertebrate lineage, two rounds of gene duplications have generated four tissue-specific vertebrate isoforms. Type I functional divergence (altered selective constraint) is significant after the first round of gene duplication. This divergence occurred mainly in the JH1 domain, but not the case in the JH2
domain, indicating that the JHI domain probably plays a major role in functional specification among tissue-specific isoforms.

Since our analysis is based on the amino acid alignment, reliability of the multiple-alignment is crucial for our interpretation. It has been indicated that the coefficient of Type I functional divergence (θ) can be overestimated by the misaligned sequences (Gu 1999, 2001). Indeed, we found a larger estimate of θ (θ = 0.615) between JHI and JH2 for the unedited Pfam domain HMM alignment, which includes several apparent misaligned sites. After making manual corrections at these sites according to the consensus sequence of protein kinase domains that have been confirmed by the secondary structure, the estimate of θ is reduced to 0.439. Furthermore, we tested the effect of the alignment by shifting the alignment window between JH1 and JH2 from -10 to +10 positions. Figure 6 shows that all estimates of θ that resulted from the shifted alignment are extremely high. Our study suggests that a reasonably good alignment is a prerequisite for estimating the level of functional divergence.

Gu’s (1999) method applies only to the Type I functional divergence that resulted in site-specific altered selective constraints among homologous genes within a gene family. For

**Figure 6.** The correlation of θ value with sequence alignment. Y-axis represents different θ values corresponding to the alignment by sliding positions (-10 to +10) relative to the manually adjusted alignment of kinase domains in Jaks.
example, the amino acid site 137 shows a typical Type I pattern, i.e. invariant tyrosine in all
tandem kinase domains, but highly variable in pseudokinase domains. It should be noted that
it is only one of many approaches to detect functional divergence from the evolutionary
perspective. As indicated by many authors (e.g., Casari et al. 1995; Livingstone et al. 1996;
Gaucher et al. 2001) functional divergence may result in dramatic change in amino acid
properties but not in selective constraints, which is also called Type II functional divergence
(Gu 1999). For instance, cluster-specific sites (or diagnosis site) usually refer to those sites
that, though highly conserved within gene clusters, consist of amino acids that are
dramatically different (e.g., positive-charged vs. negative-charged) between clusters. Another
well-cited example is that site-by-site dependence may be related to the site interaction in the
protein’s 3D structure (Pollock et al. 1999). A sophisticated model that includes all these
evolutionary aspects of functional divergence of proteins is certainly desirable. However,
because of large volume of unknown parameters, its efficiency in practice often becomes the
major concern. Therefore, a specific model is useful to test the statistical significance of a
specific evolutionary perspective. Combined with previous studies (e.g., Gu 1999, Gaucher et
al. 2001; Wang and Gu 2001), our analysis has indicated that site-specific altered selective
constraint occurs after gene duplications as a result of functional divergence in protein
evolution.

Acknowledgement: We are grateful to Dr. William Nordstrom and Peter Vedell for
constructive comments. This study was supported by NIH grant RO1 GM62118 (to X. G). X.
Gu is the DuPont Young Professor.

References
Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
Nucleic Acids Res. 25: 3389-3402
kinases and their role in growth and disease. Life Sciences 64: 2173-2186
Binari R, Perrimon N (1994) Stripe-specific regulation of pair-rule genes by hopscotch, a
putative Jak family tyrosine kinase in Drosophila. Genes Dev. 8: 300-312


JAK/STAT signal transduction pathway in response to interleukin-6. *EMBO J.* **14:** 1421–1429


CHAPTER IV: EVOLUTIONARY ANALYSIS OF FUNCTIONAL DIVERGENCE IN TGF-β SIGNALING PATHWAY

A paper published in the Information Sciences\(^1\)

Jianying Gu\(^2\), Xun Gu\(^2, 3\)

Abstract

TGF-β cytokines and their interacting receptors form an intricate signaling network and play important roles in cell differentiation and development. Members in gene families involved in this complicated pathways are usually conserved in sequence but distinct in temporal and tissue-specific expression. Starting from an evolutionary perspective, we have statistically tested the functional divergence in major pathway components, targeted the critical amino acid residues responsible for functional divergence. Our results show that altered functional constraints is a common pattern in the evolution of TGF-β pathway. Moreover, we have found the correlation between structural divergence and functional divergence in the extracellular ligand-binding domain of Type II receptors. This study helps understand the complex intrinsic mechanisms of signaling.

1. Introduction

The transforming growth factor β (TGF-β) family plays indispensable roles in tumor suppression, cell proliferation, cell differentiation, tissue morphogenesis, lineage determination, cell motility and apoptosis [1]. TGF-β and related factors regulate gene expression by bringing together Type I and Type II receptors of serine/threonine protein kinases. Interestingly, TGF-β and its downstream receptors are expressed in a well-coordinated temporal and spatial pattern. This typical ligand-receptor signaling mode provides a basis for understanding the complex regulation network in higher animal kingdom. Noticeably, considerable sequence similarity has been found across species in gene

\(^1\) Reprinted with permission of Information Sciences, 2002, 145(3-4), 195-204.
\(^2\) Graduate student and Associate Professor, respectively, Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University.
\(^3\) Author for correspondence
families involved in TGF-β pathway, which implies that certain level of functional constraint exerts during evolution [2]. Thus, an interesting question is addressed: what lies beneath the similar member genes to make the differential expression pattern of each component in this array of signaling? In this study, we shall perform comprehensive statistical analysis to investigate such functional divergence among conserved families of TGF-β pathway components from an evolutionary perspective.

2. Methods

2.1. Sequences

Amino acid sequences of TGF-β family, Type I receptor family and Type II receptor family were searched against the Non-redundant Protein databases at NCBI by the PSI-BLAST. Partial (<80% of full length) and hypothetical sequences were removed from further analysis. The accession numbers are available upon request.

2.2. Multiple Alignments and Phylogenetic Analysis

The multiple alignments were obtained by Clustal X. Phylogenetic trees were inferred by the neighbor-joining (NJ) method using MEGA2.0 (http://www.megasoftware.net/). PAUP4.0 and PHYLIP were used to examine the robustness of tree topology.

2.3. Statistical Analysis on Functional divergence - DIVERGE software

It is well recognized that gene duplication is one of major resources for functional innovation [3, 4]. The functional divergence refers to a particular type of functional innovation after gene duplication that resulting in altered functional constraints (i.e., different evolutionary rates) between two duplicate genes [5, 6].

Under a two-state probabilistic model, for each duplicate cluster, the evolutionary rate at some sites may differ from the ancestor, these sites are called F1-sites (functional divergence-related); otherwise they are called F0-sites. Rate difference between duplicates is observed only if an F1 site is present in at least one cluster: a so-called S1 status. We define the coefficient of type I functional divergence (θ) as the probability of a site being involved in functional divergence in at least one gene cluster, i.e., θ = P(S1). In contrast, a site being F0 in both clusters holds a S0 status (i.e., the evolutionary rate of each duplicate remains the same.
as the ancestral state). Maximum likelihood estimate of $\theta$ can be obtained when the phylogeny of the gene family is given [6].

If $\theta$ is statistically greater than 0, it indicates that functional divergence may have occurred after the gene duplication. To target the critical amino acid residues responsible for functional divergence, a posterior analysis is developed to score the probability for each site being functional divergence related, i.e. $P(S_t|X)$, where $X$ is the observed amino acid configuration.

The detailed statistical framework is discussed in Gu [5, 6], and Wang & Gu [7]. The computational package DIVERGE can be downloaded at http://xgu1.zool.iastate.edu/.

2.4. Dating gene duplication events

Linearized NJ trees were used to convert the (average) distance to the geological time scale. To date the gene duplication events, several speciation time points were used for calibration: primate-rodent (80 million years ago, mya), mammal-bird (310 mya), mammal-amphibian (350 mya), tetrapod-teleost (430 mya), vertebrate-Drosophila (830 mya) [8].

3. Results and Discussion

3.1. A scenario of TGF-β pathway: Ligand-Receptor Signaling

The most striking feature of TGF-β pathway is its fine modulation on a series of different factors through a tight ligand-receptor interaction. The ligands, TGF-β family members bind to receptors (Type II and Type I, consecutively) that possess intrinsic serine/threonine kinase activity. Upon ligand binding the receptors form an active heterotetrameric complex which initiates down-stream intracellular signaling [1].

3.2. Functional Divergence in Ligands

A variety of TGF-β ligands with distinct signals have been identified [2]. They belong to three major families: BMP/GDF (bone morphogenetic proteins/growth and differentiation factors), TGF-β, Activin/Inhibin, each of which has undergone substantial gene duplications. Table 1 shows the ML estimates of $\theta$ values (coefficients of functional divergence) for pairwise comparison of major member genes. Among 10 estimates, all except one are significantly greater than 0, suggesting that altered functional constraint after gene
Table 1. The pairwise coefficient of functional divergence ($\theta$) in TGF-β superfamily.

<table>
<thead>
<tr>
<th>subfamily</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>$\theta \pm$ s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGF-β</td>
<td>TGF-β1</td>
<td>TGF-β2</td>
<td>0.338 ± 0.123</td>
</tr>
<tr>
<td>TGF-β1</td>
<td>TGF-β3</td>
<td></td>
<td>0.274 ± 0.098</td>
</tr>
<tr>
<td>TGF-β2</td>
<td>TGF-β3</td>
<td></td>
<td>0.619 ± 0.126</td>
</tr>
<tr>
<td>Activin</td>
<td>Activin βA</td>
<td>Activin βB</td>
<td>0.164 ± 0.088</td>
</tr>
<tr>
<td></td>
<td>Activin βA</td>
<td>Activin βC/E</td>
<td>0.287 ± 0.138</td>
</tr>
<tr>
<td></td>
<td>Activin βB</td>
<td>Activin βC/E</td>
<td>0.115 ± 0.105*</td>
</tr>
<tr>
<td>BMP</td>
<td>BMP2</td>
<td>BMP4</td>
<td>0.250 ± 0.066</td>
</tr>
<tr>
<td></td>
<td>BMP2/4</td>
<td>BMP7</td>
<td>0.482 ± 0.074</td>
</tr>
<tr>
<td></td>
<td>BMP2/4</td>
<td>BMP9/10</td>
<td>0.330 ± 0.069</td>
</tr>
<tr>
<td></td>
<td>BMP7</td>
<td>BMP9/10</td>
<td>0.513 ± 0.111</td>
</tr>
</tbody>
</table>

* we fail to reject the null hypothesis: $\theta = 0$.

duplication is a general pattern to achieve novel functionality. In particularly, we have surveyed the pattern of functional divergence within BMP2/4 and TGF-βs.

3.2.1. BMP2/4

Both BMP2 and BMP4 are important factors for bone morphogenesis in vertebrate [9]. Tissue-specific expression has been found: in the early limb bud, BMP-4 is found in the mesenchyme and in the ectodermal ridge at the distal end, whereas BMP-2 is only found in the ridge, but not in the mesenchyme [10]. Two closely related clades in the NJ tree suggests that BMP2 and BMP4 are likely to be generated by a gene duplication event in the very early vertebrate stage, i.e., after the emergence of amphioxus and before the divergence of tetrapods and bony fishes, approximately 588 mya (Fig. 1A). Extensive studies have accumulated evidence that this short time span (430–700 mya) is important for the emergence of major molecular pathways such as signal transduction, and tissue-specific isoforms in vertebrates [11].

The overall functional divergence between BMP2 and BMP4 is indexed by the coefficient of functional divergence $\theta$ (0.250 ± 0.066). We statistically rejected the null hypothesis $H_0: \theta = 0$, and proved that altered functional constraints occurred in at least some of amino acid residues. By using the posterior analysis, we predicted that among total of aligned 338 sites, 26 sites are critical for the mutual functional divergence, with the cut-off
Fig. 1 Phylogeny and functional divergence in BMP2/4. (A) Phylogenetic tree of BMP2/4 by Neighbor-joining method. Bootstrapping values more than 50% were presented. (B) Sites predicted to be critical for functional divergence between BMP2 and BMP4.

value $P(S_i|X) = 0.51$ (as these sites are moved from the alignment, $\theta$ drops to 0). They can be grouped into two categories: Category I: high conservation observed in BMP2 sites, while considerable variation in BMP4 sites; Category II, in contrast, almost invariant residues in BMP4, but variable in BMP2 (Fig 1B). Both categories show the typical pattern of functional divergence: altered evolutionary rates and resulting functional divergence after gene duplication, in which one duplicate copy retains conserved functions and accepts most of selection pressure, while the other copy is free to accumulate amino acid changes for functional novelty.

3.2.2. TGF-β

TGF-β, whose major functional role is in cell differentiation and tissue repair, is composed of three endogeneous growth factors: β1, β2, and β3. Despite the sequence and structural similarities [12, 13], these three isoforms have been shown apparent differences in activity such as ligand binding affinity [14].
Fig. 2 Phylogeny and functional divergence in TGF-β ligand family. (A) Neighbor-joining tree of ligands TGF-βs. Bootstrapping values more than 50% were presented. (B) site-specific profiles for pairwise comparison of three TGF-β isoforms. P(S|X) is the posterior probability for a site being functional-divergence related.

Phylogenetic analysis shows that three TGF-β isoforms are monophyletic (Fig. 2A). Two rounds of gene duplication may have occurred to give rise to three paralogous clusters after the divergence of tetrapods and fishes: the first round generates β1 and the common ancestor of β2 and β3, then the second round produces β2 and β3. This provides evidence that TGF-β ligand may belong to a large group of cytokines that emerged after large-scale or genome-wide duplications [11].

The functional divergence analysis reveals that the coefficients of functional divergence (θ) for pair-wise comparison of TGF-βs are significant, suggesting detectable functional alteration occurred in the lineage leading to them (Table 1). Furthermore, posterior statistics implies that strong variation in contribution of each site to the overall...
functional divergence. Figure 2B presents respective site-specific profiles which have the common pattern. The baseline of posterior probability for a site being functional divergence related is about 0.2-0.3, implying that most of amino acid residues do not have significant effects on long-term functional divergence; while several peaks show that only a small portion of sites are mainly responsible for functional diversification in two paralogs. This preliminary targeting strategy along with other available computational approaches provide a start point for future experimental detection of 'hot-spots' in tissue-specific isoforms [15].

3.3. Functional Divergence in Receptors

Two types of receptors, Type II Receptor and Type I Receptor, have been identified to bind specific TGF-β ligands. They belong to a large serine/threonine kinase superfamily [1].

3.3.1. Type II receptor: structural divergence → functional divergence?

Type II receptor family is divided into five subfamilies: Activin Receptor II (ActR-IIA), ActR-IIB, TGFβ-IIR, BMPR-II, and AMHR. The existence of invertebrate homologs (Drosophila, C.elegans, and sponge) suggests the ancient origin of this family: ActR-IIA and IIB form a more closely related cluster, while others might have diverged earlier (Fig. 3).

It has been recognized that structural difference may result in the functional significance [16]. Our recent study in apoptotic caspases has shown the structural basis of functional divergence [7]. Analogously, we found that the functional divergence between two Type II receptor subfamilies, ActR-IIA and AMHR, may directly associate with their structural alteration. ActR-IIA binds specifically to Activin, while AMHR only recognizes MIS. The 3-D structure of ActR-IIA has been revealed and its geometrical difference with AMHR has been discussed [17]. We obtained the coefficient of functional divergence between them (θ = 0.709 ± 0.232). Moreover, 33 sites are predicted critical for their functional divergence and then mapped onto the 3-D structure of ActR-IIA. Figure 4 shows the alignment of extracellular ligand binding domain. Notice that 13 predicted residues lie in important secondary structures (α helices, β sheets or strands). Similar to the pattern in BMP2/4 subfamilies, these sites are only conserved in one cluster, either ActR-IIA or AMHR, but variant in the other. Although the intrinsic mechanism is unclear, the sequence
alteration may result in structural changes and further functional changes. Interestingly, the sites which believed to play important role in ligand recognition: site 106, site 108 and site 109 (as shown in Fig. 4) show a noticeable pattern. In ActR-IIA those positions hold highly conserved Asp, Asp and Lys, while in AmhR are Arg, Lys and Tyr. The amino acid configurations are invariable in both clusters but whose biochemical properties are very different, which suggested that these sites might determine the ligand specificity of different receptors.

3.3.2. Type I receptor

Based on the identity of physiological ligands, Type I receptor is divided into 7 major subfamilies: TβR-I, ActR-Iβ, ALK7, BMPR-IB, BMPR-1A, ALK1, and ActR-I. NJ tree shows that they fall into separate clades with high bootstrap support (>95%) (Fig. 5). Based on kinase domain similarity and signaling activities, they can be clustered into three groups.
Fig. 4 Sequence alignment of extracellular ligand-binding domain of ActR II-A and AmhR. Secondary structure elements refer to the sequence of human ActR II-A. Predicted critical residues for functional divergence are listed (site number corresponds to position in sequence alignment of full length genes).

One group includes TβR-I, ActR-Iβ and ALK7; another includes BMPR-IB and BMPR-IA; and the third includes ALK1, and ActR-I.

The functional divergence analysis of pair-wise comparisons of Type I receptor showed the statistically non-zero coefficient of functional divergence (θ) (data not shown). This analogous result to ligands and Type II receptor suggested that altered functional constraint after gene duplication is a general pattern at different levels of signaling pathway.
Fig. 5 NJ tree of Type I TGF-β receptor.

4. Conclusions

In this paper, we have explored the functional divergence in TGF-β signaling. Our findings are: (1) The altered functional constraint is a common pattern in both ligand and receptor evolution; (2) Gene duplication might be an evolutionary force for functional divergence, the early vertebrate stage might be an important time for the development of tissue specificity and signaling pathways (3) There may be correlation between structural and functional divergence.

Acknowledgement

This study was supported by NIH grant RO1 GM62118 (to X.Gu).

References


CHAPTER V: EVOLUTIONARY PATTERN OF PROTEIN KINASE EXPANSION IN HUMAN GENOME

A paper to be submitted\textsuperscript{1}

Jianying Gu\textsuperscript{2}, Xun Gu\textsuperscript{2,3}

Abstract

Kinases catalyze the reversible protein phosphorylation and play a central role in fundamental biological functions and in advanced cell-cell co-regulation. A comprehensive set of kinases (kinome), characterized by a conserved catalytic kinase domain, has recently been identified in human genome through exhaustive bioinformatic search. We conducted phylogenetic analysis on major kinase gene families. The age distribution of kinase gene families showed that ancient continuous domain shufflings (or duplications) during the time period from early stage of eukaryotes to metazoan evolution have built up the major kinase-related animal specific signal transduction network. The following large scale gene duplications in the early stage of vertebrate have contributed to the tissue-specific signal transduction. Moreover, several kinase pseudogenes are likely to be generated through either segmental duplication or retrotransposition very recently.

Introduction

Reversible protein phosphorylation plays a central role in regulating many functions of eukaryotes such as cell cycle control, gene transcription and protein translation, apoptosis, cell differentiation, cell-cell communication, and to mediate complex interactions with the external environment. Moreover, aberrant protein phosphorylation is commonly related with cancer and other human diseases. A comprehensive knowledge of the key actors that regulate these functions can provide basis for better understanding of underlying mechanism for cell life and novel drug discovery strategies.

\textsuperscript{1} manuscript to be submitted.
\textsuperscript{2} Graduate student and Associate Professor, respectively, Department of Zoology and Genetics, Center for Bioinformatics and Biological Statistics, Iowa State University.
\textsuperscript{3} Author for correspondence
The completion of human genome sequence provides us an opportunity to decipher the molecular nature of signal transduction machinery in human. A total of 478 protein kinases have been identified in human genome, constituting about 1.7% of all human genes (Manning et al., 2002). All of known bona fide protein kinases share a conserved catalytic kinase domain of ~ 270 amino acids formed a superfamily - eukaryotic protein kinase (ePK). Two distinct major families, protein tyrosine kinases (PTKs) and protein serine/threonine kinases (PSKs), are defined based on comparisons of kinase domains. Another set of 40 atypical protein kinases (aPK) have also been identified in human genome, with demonstrated kinase activity despite the lack of sequence similarity to ePK domain.

It has been widely accepted that the duplications of individual gene, chromosome segment, or whole genome provide the primary material for origin of functional novelties (Ohno, 1970). Recently, we have demonstrated that the age distribution of human gene families can be described as an ancient component, a peak in the early stage of vertebrates (wave II), and a recent peak during mammalian radiation (wave I) (Gu et al., 2002). We have further speculated that the major signal transduction pathways emerged through those duplication events. It is indeed intriguing to study the functional divergence and evolutionary patterns of the entire set of kinase (kinome, defined by Manning et al. 2002). Here we report the age distribution of human kinases and pseudogenes and show a complex history of domain shufflings, gene duplications, and gene losses in kinome expansion.

Data and Methods

Data collection

In order to identify the homologues of 518 human protein kinases in vertebrates such as other mammals, birds, frogs, and fishes, we searched HOVERGEN protein database (Homologous Vertebrate Genes Database, http://pbil.univ-lyon1.fr/databases/hovergen.html, release 42 updated on 26 April, 2002) using BLASTP, followed by a pairwise BLAST to removed the redundant sequences. Invertebrate homologous sequences were obtained from http://www.kinase.com, and used as outgroup for each gene family. Domain organization of each kinase gene family was revealed by searching against the Pfam domain database (http://pfam.wustl.edu/).
The orthologous mouse DNA sequence for human kinase pseudogenes were obtained by BLASTN searching against ENSEMBLE mouse genome (http://www.ensemble.org). Three MARK pseudogenes (MARKps10, 13, 20) were not analyzed due to short length (< 200 nucleotide acid). With a cutoff E value 1e-20, two pseudogene SAKps, SGK050ps has NO mouse hits, which will be excluded from further analysis.

**Multiple alignments and phylogenetic inference**

Multiple sequence alignment of each gene family was obtained by Clustal X (Thomson et al., 1997), followed by manual editing. Phylogenetic trees were inferred by the Neighbor-joining method (Saitou and Nei, 1987) using MEGA2.0 (http://megasoftware.net). Maximum Parsimony (as implemented in PAUP4.0), and Maximum Likelihood (as implemented in PHYLP) were used to examine whether the inferred phylogeny is sensitive to any tree-making method. The bootstrap resampling with 1,000 pseudo-replicates was carried out to assess support for each individual branch.

**Inference of duplication events**

To identify the underlying evolutionary mechanism (i.e., domain or gene duplication) of each duplication event, we have examined the domain structure of each human protein kinase as follows. First, homologous kinase sequences in other vertebrate species (e.g., other mammals, birds, frogs, and fish) were identified by blasting the 518 human kinase protein sequences against the HOVERGEN protein database. A total of 221 HOVERGEN families were retrieved and further analyzed. Second, since the HOVERGEN family is defined solely based on the overall sequence similarity and may not reflect functionality, we adopted the hierarchical kinome classification scheme proposed by Manning et al. (2002) which incorporates sequence inference, structure comparison, and known biological functions. Third, for multi-domain kinase genes, in addition to the kinase domain, additional domain structure was revealed by using the HMM and profile-searching in the Pfam database (http://pfam.wustl.edu/). Finally, combining the information from all the previous analysis, we classified kinases with similar multi-domain structure into 256 gene families. Here, we define the duplication event occurred within each gene family as gene duplication (Fig. 1), and that occurred among gene families as domain duplication (Fig. 2). The remaining single-
domain kinases (e.g., CDK family) that lack of hierarchical domain structure were analyzed separately.

![Phylogenetic Tree]

**Fig. 1.** Molecular Evolutionary analysis of the Rac-beta serine/threonine protein kinase (AKT) gene family. In vertebrate, there are three member genes: AKT1, AKT2 and AKT3. (a) The phylogenetic tree inferred by the neighbor-joining method with Poisson distance. Bootstrapping values of more than 50% were presented. $T_1$ and $T_2$ are the time points of the first and second gene duplications, respectively. (b) The linearized neighbor-joining tree used in converting evolutionary distance to the (relative) molecular time scale. By the nearest-neighbor clock, we estimated $T_1 = 478$ Mya and $T_2 = 522$ Mya.
Fig. 2. Phylogenetic relationship and the domain structures of protein tyrosine kinases AXL and ROR gene family.

A linearized neighbor-joining tree is efficient to convert the (average) Poisson distance of protein sequences to the molecular time scale. Here primate-rodent (80 mya), mammal-bird (310 mya), mammal-amphibian (350 mya), tetrapod-teleost (430 mya) and vertebrate-Drosophila split (830 mya) were used as calibration in our study (Wang and Gu, 2001). Two steps were taken to adopt a nearest-neighbor clock (Gu et al., 2002). First, under the tree, the phylogenetic interval of a duplicate event is determined by the nearest speciation event(s). It can be two sides (before, after), e.g., (teleost-tetrapod, Drosophila-
vertebrate), or one side, e.g., (teleost-tetrapod, \( \infty \)) or \((M, \text{teleost-tetrapod})\) for before or after the speciation, respectively, where \( M \) (the mammalian radiation) is for the lowest side. Note that the phylogenetic interval is robust against the differential selective constrains in genes and lineages. Second, the age of duplication event is estimated based on the nearest speciation event(s). In practice, if the "before"-side is used, an average over two lineages is taken when they have the same nearest speciation. If the phylogenetic interval is two-sided, one may use two points for dating (Fig. 1).

To date the ancient duplication events occurred in early Metazoan evolution, a hierarchical nearest neighbor clock is applied. The dates of the nearest duplication events inferred by the previous nearest-neighbor clock or nearest speciation points (chosen based on its applicability) were used to infer the ancient time point of duplication events.

**Analysis of kinase pseudogenes**

There are two major types of pseudogenes: processed pseudogene and duplicated pseudogenes. Processed pseudogenes result from retrotransposition, that is, reverse-transcription of mRNA transcript followed by integration into genomic DNA. Duplicated pseudogenes arise from duplication of genomic DNA, after which a duplicate gene copy became a pseudogene. Neighbor-joining tree was inferred from the multiple alignment of DNA sequences of kinase pseudogene, its parent human kinase gene, homologous mouse gene, and homologs in other vertebrate species (Fig. 3). The branch length in NJ tree represents the number of nucleotide substitution per site (under Kimura 2 parameter model) along that lineage. Due to relaxation of selection constraint, fast evolution occurred along the pseudogene lineage. Thus we estimate the time of the events to generate pseudogene assuming a local molecular clock between the functional human/mouse kinase genes.

**Results**

*Protein kinases: An evolutionary overview*

We have reconstructed phylogenetic trees for 518 human protein kinases and 106 human pseudogenes, which include their homologous genes in vertebrate (mammals, birds, frogs, and fishes), invertebrate (largely fruitfly and worm), and other eukaryotes. Our phylogenetic analysis is generally consistent with the human kinome classification (Manning
Fig. 3. Molecular phylogeny of the P21-activated kinase (PAKA) gene and pseudogene. The branch length represents the number of nucleotide substitution per site under a Kimura two-parameter model. The human-mouse split $T = 80$ Mya is used to estimate $T_1$ to be 18 Mya, the time point to generate PAK pseudogene.

et al., 2002). We have confirmed that most protein kinase families had existed before the origin of vertebrate and most subfamilies have emerged in the early stage of vertebrates (Manning et al., 2002b; Miyata and Suga, 2001). Protein tyrosine kinase (PTKs), one major animal-specific subclass that includes 29 families, is likely monophyletic, derived from a precursor of protein serine-threonine kinase (PSK) during the course of animal evolution, while the major types of PSKs had existed in the early stage of eukaryotes (Hanks et al. 1988; Hanks and Hunter, 1995; Hunter et al. 1995).

Comparison of the multiple domain structure of protein kinases showed that the kinase domain is the only conserved region among most protein kinase families, indicating a reminiscent important role of domain shuffling/duplication on the emergence of major signal transduction pathways featured by different protein kinases. In contrast, the multiple domain structure of most kinase families remains conserved among member genes, suggesting the impact of gene or genome duplications on protein kinase expanding toward development/tissue specificity. Moreover, only a very few (~2) human protein kinases are
recently duplicated (after human/mouse split), in spite of recently finding that a significant portion of human genome has been duplicated within 40 million years (Lander et al., 2001).

To further characterize the relationship between underlying evolutionary mechanisms, evolutionary time periods, and the evolution of signal-transduction pathway, we have developed a robust procedure to infer the events of domain shuffling, gene duplication, or pseudogene duplication (see Methods). The 524 duplication events identified all occurred in the ancestral lineage of humans. Among them, 95 are pseudogene duplication/retrotransposition events, 222 are gene duplications, and 187 are domain duplication events. We are not able to classify the rest 20 duplication events that occurred among single-domain kinases as domain or gene duplications.

Remarkably, the age distribution of protein kinases and pseudogenes has recaptured the unique evolutionary features of human gene families as first described by Gu et al. (2002). Clearly, the age distribution of human kinome shows a pattern characterized by three modules (I, II, and III) (Fig. 4). Module I coincides with the processed pseudogene explosion (Zhang et al., 2002) or abundant recent tandem or segmental duplications (Eilcher, 2001). Module II indicates kinase gene family expansion in the early stage of vertebrate evolution by genome (large-scale) duplication (Gu et al., 2002; McLysaght et al., 2002). The ancient module (III) includes duplication events that took place during and before metazoan evolution.

**Ancient components: domain shuffling/duplication**

Domain shuffling (or duplication), which can occur by transposition of gene fragments or nonhomologous recombination, is thought to have been a major mechanism in the evolution of new multiple-domain proteins (Doolittle, 1995). Like many other molecules involved in signal-transduction, most protein kinases are composed of multiple domains with discrete structural and specific functions. Separate phylogenetic analysis of protein kinase domain and their associated non-catalytic domains (SH2, SH3, PHD, etc.) has confirmed that the kinase superfamily was indeed generated by shuffling of domains from a few more ancient kinases (Krupa and Srinivasan, 2002).

It is known that time estimation of ancient evolutionary events (either duplication or speciation) may have broad sampling variance or bias (Gu, 1998). Nevertheless, the age
Fig. 4. Age distribution of human kinases. The molecular time scale is measured as Myr ago (Myr). Each bin for the histogram is 50 Myr, except for the most ancient one (>1,500 Myr ago). 1 Byr, 1 billion years (1,000 Myr).
distribution of domain shuffling events in protein kinases shows that the majority of domain duplication events happened during very ancient time period. As shown in Fig. 4, no domain shuffling event is found after 550 mya, and the majority had occurred during the time period from early stage of eukaryotes to metazoan evolution. Conserved motifs shared with eukaryotes protein kinases were identified in bacteria, as well as in archaea, which inferred the existence of an ancestral protein kinase prior to the divergence of eukaryotes, bacteria and archaea (Leonard et al, 2002). However, the biological reason why domain shuffling virtually ceased after the origin of vertebrates remains unclear.

On the basis of phylogenetic analysis of eight animal-specific gene families including sponge protein tyrosine kinases, Suga et al. (2001) suggested that most domain shuffling events were going back to dates before the parazoan-sumetazoan split (~ 940 mya), the earliest divergence among extant animal phyla. Our comparative analysis, however, does not support this notion. Rather, we provide evidence (Fig. 4) that during the time period from early stage of eukaryotes to metazoan evolution, roughly corresponding to 2,000 mya (2.0 bya, billion years ago) to 700 mya, domain shuffling in protein kinases had occurred in a fashion of continuous flux, with an approximate constant rate of $0.7 \times 10^{-9}$ event per year per genome. Before this time period (very early eukaryote stage, or prior to eukaryote-archaea split, > 2.0 bya), the rate of domain shuffling was down to ~50%, while after this time period (vertebrates), the rate is virtually zero.

**Large scale duplication of human protein kinases**

It has been proposed long time ago that genome duplications play important role in building up the complexity of human genome (Ohno, 1970). Our recent analysis of age distribution of human gene families has confirmed a rapid explosion of genes in the early stage of vertebrate (Gu et al., 2002), resulting in tissue-specific (developmental stage-specific) isoforms. Molecular phylogeny-based analysis of gene families involved in cell-cell communication and developmental control (e.g., protein tyrosine kinases, phosphatases, etc.) has shown that extensive gene duplications occurred in a period around or immediately before the divergence of tetrapods and fishes (Miyata and Suga, 2001). The recent completion of human kinome (Manning et al., 2002) provides us an opportunity to study the
evolution pattern of the complete set of human kinases, a better understanding for the whole set of human genes.

Based on sequence similarity (kinase domain similarity), domain organization and known biological functions, we identified 81 gene families with two human kinase member genes, 34 gene families with three human kinase member genes, and 14 gene families with four human kinase member genes. Our classification generated slightly different kinase gene families compared to the kinome classification scheme. No evidence supporting the 1:4 rule (i.e., there is one gene in Drosophila but four homologous genes in human) exists in the human kinome evolution. Moreover, we also examined the phylogeny of gene families with four member genes. Two types of topologies can be observed ((A, B) (C, D)) or (((A, B), C), D). There is no evidence that the topology of ((A, B), (C, D)) as expected by 2R genome duplication hypothesis occurred more frequently than the other topology.

Among 222 estimated kinase gene duplication events occurred within the gene families, 135 (~ 61%) occurred in the time period before the emergence of teleosts but after the vertebrate-amphioxus split (430-750 Mya) (module II in Figure 4). This rapid increase in the number of paralogous human kinases observed in the early stage of vertebrate evolution is as expected from the genome duplication hypothesis. Human tissue specific isoforms of protein kinases are involved extensively in signal transduction in eukaryotic cells, as well as in controlling many other cellular processes. Thus genome duplication may have major contribution to the tissue- (or developmental stage-) specific kinase isoforms after the split between vertebrate and anthropods.

Evolution of human kinase pseudogenes

For 106 human kinase pseudogenes identified by Manning et al. (2002), 7 pseudogenes are excluded from further analysis due to very short sequences. Phylogenetic analysis shows that most human kinase pseudogenes (75/99) are generated after the mammalian radiation. Because of the apparent fast evolution caused by the loss of functional constraint in the pseudogene lineage, we use some special treatments to date the duplication events which gave rise to the current pair of functional (parent) kinase gene and pseudogene (see Methods).
Generally, pseudogenes are believed to be generated through two different mechanisms: via retrotransposition resulting in an intronless processed pseudogene or via duplication of genomic DNA resulting in duplicated pseudogene. In 106 human kinase pseudogenes, 75 lack introns such that they were likely to be generated by retrotransposition of a processed transcript. It has been shown that a little more than half or even a much higher eighty percentage of pseudogenes are processed (Harrison et al., 2002; Migghell et al., 2000), so it is not surprising that the majority of human kinase pseudogenes are processed. It has been proposed that processed pseudogenes, as well as both the Alu repeats (a major class of SINEs, short interspersed elements) and LINE repeats (long interspersed elements), arose by the protein machinery encoded by LINE1 elements (Jurka, 1997; Esnault et al., 2000). The age distribution of human kinase intronless pseudogenes showed that more than half of intronless pseudogenes were generated very recently (< 30 mya) in a fairly constant rate (Fig. 5), while a considerable proportion of MARK intronless pseudogenes were generated after human-mouse split.

Fig. 5. Age distribution of human processed pseudogenes.
Twenty-nine pseudogenes were identified that had no introns. They are more likely to be generated through gene/genome duplication events, and most of them occurred on the anthropoid lineage. Recent analysis of human genome sequence has revealed the fact that about 5% of the human genome consists of interspersed segmental duplications that have arisen over the past 35 million years (Eichler 2001, etc.). The age distribution of kinase pseudogenes obviously resembles the pattern of recent segmental/tandem duplication (data not shown). At this rate, we estimated that $478 \times 5\% \approx 24$ kinase pseudogenes were generated through segmental/small-scale duplications. According to these observations, we propose a scenario for recent human protein kinase evolution: within 35 million years, segmental duplications (small-scale mode, Gu et al. 2002) provided a continuous flux for generating duplicate kinase genes, which evolved into protein kinase pseudogenes. For example, two p70 ribosomal protein S6 kinase (p70S6K) pseudogenes appears to be part of a large duplicon containing multiple duplicated genes, suggesting an intrachromosomal duplication of the p70S6K locus (Manning et al, 2002). Apparently, these duplicates finally become nonfunctional (pseudogene) because there seems no need for new protein kinases. In this regards, one may conclude that the protein kinase-related signal transduction pathway evolves slowly after the mammalian radiation, while regulatory motif changes may become the major force for functional/expressional divergence. This hypothesis should be tested by further multi-genome analysis of protein kinases.

In 106 human kinase pseudogenes, 27 are found to have matching ESTs. The age distribution of pseudogene duplications generated these “expressed” pseudogene appears random and uniform, indicating no obvious correlation between the age of pseudogene duplication and gene transcriptional regulation change. There are many possible explanations: (i) the recent produced pseudogene is still under the dynamic processing, and has not lost all its transcriptional capability; (ii) the mutations accumulated that resulted in functional loss were likely to happened in the coding region, but not in transcription regulatory region; (iii) misidentified ESTs for kinase pseudogene.

Fates of duplicate genes and the gene-loss hypothesis

Gene duplication has generally been viewed as a necessary source for the origin of novel functionality. In addition to ignite gene function innovation, gene duplication may also
provide a redundant copy such that it will lose the original function. Because the majority of
mutations are deleterious and gene duplications are generally assumed to be functionally
redundant at the time of origin, the usual fate of a duplicate-gene pair is the
nonfunctionalization of one copy (Lynch and Conery, 2000). Recently, a less-is-more
hypothesis has been proposed to emphasize the importance of loss-of-function mutations on
the recently evolved novel lineage such as the human (Olson and Varti, 2003). In several
cases, genetic loss in human lineage has been shown to cause some differences between
human and chimpanzee.

As reported in Manning et al. (2002), three pseudogenes have no obvious parent
kinase in human but have functional orthologs in rodents. Our re-examine showed that
parent kinase of polo-like kinase SGF384ps has been identified (EMBL Accession:
AK054808) in human. But for KSGC (kinase-like domain containing soluble guanylyl
cyclase) and CGD, no functional human orthologous kinase has been identified (Figure 6).
Phylogenetic analysis shows that they both belong to the retinal guanylyl cyclase family
(GTP pyrophosphate-lyase). These two pseudogenes both contain introns and conserved
canonical splice sites. Moreover, matching ESTs have been found. We inferred these two
genes lost their functions recently such that not too many mutations have been accumulated.
In rat, CGD is specifically expressed in a small, randomly dispersed subpopulation of rat
olfactory sensory neurons and plays a role in interacting with a restricted group of odors
since its expression pattern resembles the pattern of expression of the diverse seven trans-
membrane-domain odorant receptors (Fulle et al., 1995; Julifs et al., 1997). KSGC found in
rat kidney cells appear to contain the kinase-like, dimerization and catalytic domains but
containing no ligand-binding or transmembrane domains, indicating that KSGC is a
cytoplasmically localized GC (Kojima et al., 1995). The specified function and unique
domain structure of CGD and KSGC suggested that the decay of previous functional genes in
the human lineage after its split from rodents. This hypothesis could be further tested when
genome sequences from other species such as chimpanzee are available.
Fig. 6. Molecular phylogenetic analysis of retinal guanylyl cyclase. Two human kinase pseudogenes CGDps and KSGCps are indicated.
Discussion

The recently catalogued complete set of human protein kinases (kinome) allows us to test an interesting hypothesis on the evolution of protein kinases at the genome level: (i) The major kinase-related animal specific signal-transduction pathways have been generated through domain shuffling during the time period from the early stage of eukaryotes to the course of metazoan evolution; (ii) vertebrate tissue-specificity of signal-transduction is facilitated by large-scale duplication event(s) in the early stage of vertebrates, and (iii) continuous small-scale duplication events and retrotranspositions occurred recently resulted in the nonfunctional kinase pseudogenes.

Unlike the continuous mode of retrotransposition of kinase processed pseudogenes, the distribution of ribosomal protein (RP) pseudogenes shows a peak at an evolutionary age corresponding to 8% - 10% sequence divergence, whereas Alus peaks at 7% and LINE1 peak at both 4% and 21% (Zhang et al., 2002). Considering the fast evolution rate of pseudogene ($3.9 \times 10^{-9}$ substitution per year) and the relative slow evolution rate of functional gene ($1.5 \times 10^{-9}$ substitution per year) (Li, 1997), one percentage of divergence between pseudogene and functional gene was converted into approximately 2.1 Myr. Thus the peak of generation of processed RP pseudogenes appeared at 15-20 Mya and the rate of new processed RP pseudogenes generated has slow down after that. The overall activity of transposons has declined markedly over the past 35-50 Myr, with the possible exception of LINE1 (Nature human genome, 2001). Interestingly, both LINEs and processed pseudogenes are more prevalent in relatively GC-poor regions, while Alus are predominantly found in GC-rich regions (Pavlicek et al., 2001; Zhang et al., 2002). Noticeable, processed pseudogenes are not expressed, however in very rare case, transcripts of some pseudogene have been reported though the functional relevance of these pseudogene transcripts remains unclear. We observed seventeen intronless pseudogenes having matching EST, one possible explanation of this unusual high rate of transcriptional processed pseudogenes is the high sequence similarity between processed pseudogene and ESTs (both contain no intron).

Acknowledgement: We are grateful to Yufeng Wang for constructive comments. This study was supported by NIH grant ROI GM62118 (to X. G).
References


CHAPTER VI: GENERAL CONCLUSION

This study is focused on the exploitation of the functional divergence and evolutionary patterns in vertebrate kinomes and kinase-regulated signaling transduction, using a combinatorial statistical and evolutionary approach. After a general introduction in Chapter I, the major findings are reported in Chapter II to Chapter V.

Summary of Findings

In the first study (Chapter II), we investigated the natural history and evolutionary patterns in the protein tyrosine kinase (PTK) supergene family. PTK is the key mediator in cellular signaling in metazoans, and is directly associated with a variety of human diseases. Our analysis showed that site-specific shifted evolutionary rate (altered functional constraint) is a common pattern in PTK gene family evolution. Our results are summarized as follows: (1) Both large-scale and small-scale gene duplications are the major evolutionary force for generating the contemporary PTK superfamily. (2) Substantial functional divergence occurred after gene duplication(s), characterized by the significant shift in evolutionary rates. (3) Evolutionary functional divergence is correlated with the phenotypic functional divergence in paralogous genes. These results not only shed light on the role of gene duplication in the development of hierarchical PTK-mediated networks, but also provide impetus for a new approach to predict functional divergence from evolutionary changes.

In the research presented in Chapter III, we conducted a statistical and phylogenetic analysis of the Jak (Janus kinase) gene family. In addition to a typical kinase domain (JH1), the Jak family possesses a unique noncatalytic pseudokinase domain (JH2). This dual kinase domain structure provides an ideal model for testing the impact of the (internal) domain duplication on functional divergence. The impact of gene duplication, which may have given rise to four tissue-specific member genes, Jak1, Jak2, Jak3 and Tyk2 in the early stages of vertebrates, is of particular interest in this study as well. Our result shows that both domain and gene duplications are important for Jak gene family proliferation and evolution: (1) Shifted selective constraints (or shifted evolutionary rates) are statistically significant between JH1 and JH2. (2) Predicted amino acid sites responsible for difference between JH1
and JH2 by posterior analysis can be classified into two groups: very conserved in JH1 but highly variable in JH2, and *vice versa*. JH2 domain appears to have acquired some new functions that may account for its long-term existence during evolution, since a group of sites conserved in the JH2 domain are variable in the JH1 domain. (3) In the early vertebrate lineage, two rounds of gene duplications have generated four tissue-specific vertebrate isoforms: after the (first) gene duplication, site-specific rate shifts between Jak2/Jak3 and Jak1/Tyk are significant. This divergence took place mainly in the JH1, but not in the JH2 domain, indicating that the JH1 domain probably plays a relatively important role in functional specification among tissue-specific isoforms.

In Chapter IV, we have explored the functional divergence in major pathway components of kinase-mediated TGF-β signaling pathways from an evolutionary perspective. Members in gene families involved in this complicated pathway are usually conserved in sequence but distinct in temporal and tissue-specific expression. Our major findings include: (1) The altered functional constraint is a common pattern in both ligand and receptor evolution in TGF-β signaling; (2) Gene duplication might be an evolutionary force for functional divergence; (3) The early vertebrate stage might be an important time span for the development of tissue specificity and signaling pathways; and (4) The correlation between structural and functional divergence seems to be evident.

After the analysis of an individual kinase gene family (Jak), protein tyrosine kinase superfamily, and a kinase mediated signaling transduction pathway (TGF-β), we have attempted to globally explore the functional divergence in the whole set of vertebrate kinases (kinome) (Chapter V). The age distribution of kinase gene families showed that (1) The major kinase-related animal specific signal-transduction pathways have been generated through an ancient continuous domain shufflings (or duplications) during the time period from early stage of eukaryotes to metazoa evolution; (2) Vertebrate tissue-specificity of signal-transduction is facilitated by large-scale duplication event(s) in the early stage of vertebrates; and (3) The kinase pseudogenes are generated through either segmental duplication or retrotransposition very recently. These findings are essentially congruent with our earlier report that both large scale and small scale gene duplications have significant contribution during vertebrate evolution.
Ongoing and Future work

My long-term research goal is to integrate multiplayer genome information to study evolutionary/comparative genomics and computational biology. The accomplished researches reported here represent my effort to systematically study the evolutionary mechanisms in complex biosystems, using kinase (kinome) as a paradigm. In the short term, I hope to develop research programs to:

- Exploit the temporal-spatial expression profile of kinome, with a focus on the statistical tests on publicly available microarray data (Gu and Gu 2003).
- Further investigate the kinome proliferation on functional innovations from tissue-specificity to molecular pathways, and to reveal the joint distribution of age and chromosome location of duplication genes.
- Extend the study of functional divergence to other important families in kinase signaling network. The target families include transcription factors and G-protein coupled receptors.
ACKNOWLEDGEMENTS

A journey is easier when you travel together. Interdependence is certainly more valuable than independence. This thesis is the result of five years of work whereby I have been accompanied and supported by many people. I am glad to have the opportunity to express my gratitude for all of them.

I am deeply indebted to my major advisor, Dr. Xun Gu, whose insight, stimulating suggestions and encouragement helped me in all the time of research of this thesis. His enthusiasm and integral view on science has made a deep impression on me. I would also like to thank other committee members Dr. Dan Nettleton, Dr. Linda Ambrosio, Dr. Gavin Naylor and Dr. Daniel Voytas for their support during my Ph.D., especially to Dr. Dan Nettleton for his guidance and advice during my study for M.S in statistics.

My research work is also strongly supported by the members of Computational Genomics and Evolution Lab. I want to thank them for all their help, support, interest and valuable hints. Especially I am obliged to Dr. Yufeng Wang. Our discussions of coursework, research and life-in-general were always a welcome addition.

I would like to give my special thanks to my family. They have contributed so much to my life that I could never acknowledge it all in this limited space.
VITA

NAME OF AUTHOR: Jianying Gu

DEGREES AWARDED:
  B.S. in Microbiology, Fudan University, 1995
  M.S. in Molecular Genetics, Shanghai Institute of Plant Physiology, 1998
  M.S. in Statistics, Iowa State University, 2002

HONORS AND AWARDS:
  James Cornette Research Excellence Award in Bioinformatics, Iowa State University, 2002

PROFESSIONAL PUBLICATIONS: