2007

Functional divergence after gene duplication and sequence-structure relationship: a case-study of G-protein alpha subunits

Ying Zheng
Iowa State University

Functional divergence after gene duplication and sequence-structure relationship: A case-study of G-protein alpha subunits

by

Ying Zheng

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Xun Gu, Co-Major Professor
Karin Dorman, Co-Major Professor
Xiaoqiu Huang

Iowa State University

Ames, Iowa

2007

TABLE OF CONTENTS

ABSTRACT

In this study, I use animal G-protein alpha subunit family as an example to illustrate a comprehensive analytical pipeline for detecting different types of functional divergence of protein families, which is phylogeny-dependent, combined with ancestral sequence inference and available protein structure information. In particular, I focus on (*i*) Type-I functional divergence, or site-specific rate shift, as typically exemplified by amino acid residue highly conserved in a subset of homologous genes but highly variable in a different subset of homologous genes, and (*ii*) Type-II functional divergence, or the shift of cluster-specific amino acid property, as exemplified by a radical shift of amino acid property between duplicate genes, which is otherwise evolutionally conserved. We utilized the software DIVERGE2 to carry out these analyses. In the case of G-protein alpha subunit gene family, I have tested the significance of functional divergence between subfamily Gq and Gs, and predicted the candidate amino acid residues related to either Type-I or Type-II functional divergence. Then, the inferred ancestral sequences and current amino acid configuration of these candidate sites were combined with phylogenies to explore the trends of functional divergence. Finally, these predicted residues are mapped to the protein structures to test whether these residues may have structures or solvent accessibility preferences.

CHAPTER I. GENERAL INTRODUCTION

1.1 Gene duplication and gene families

A gene family is a set of genes evolved from a common ancestral gene. They generally share similarity whereas with divergence in sequence, structure/function. Gene duplications and domain duplication/shuffling, have been considered to provide the major source of the innovation and complexity of gene families (Ohno, 1970; Doolittel, 1995; Sidow, 1996; Henikoff et al., 1997; Li et al, 2001).

1.2 Functional divergence of gene families

Understanding functional divergence after gene duplication is one of the major goals in functional genomics (Henikoff et al. 1997; Bork and Koonin 1998). Under the framework of phylogenomic annotation of gene function (Golding and Dean 1998; Eisen and Fraser 2003), the importance of gene 'function' can be measured quantitatively in terms of the functional constraints of the protein sequence (Kimura 1983). As an amino acid residue is said to be functionally important if it is evolutionarily conserved, it has been recognized recently that change of the evolutionary conservation at a particular residue may indicate its involvement of functional divergence (Lichtarge et al. 1996; Gu 1999). Following this idea, many research groups including ours have developed statistical methods for testing and predicting functional divergence after gene duplication (e.g., Lichtarge et al. 1996; Gu 1999, 2001, 2006; Landgraf et al. 2001; Knudsen and Miyamoto 2001; Lopez et al. 1999; Gribaldo et al. 2003; Madabushi et al. 2004; Gao et al. 2005). Based on these

methods, many case studies have shown the association between sequence and function/structure divergence (e.g., Gaucher et al. 2002a; Landgraf et al. 2001; Wang and Gu 2001; Jordan et al. 2001; Gribaldo et al. 2003; Gao et al. 2005; Rastogi and Liberles 2005; Zhou et al. 2006).

1.2.1 Types of functional divergence in protein sequence evolution

From the view of molecular evolution, an amino acid residue is said to be functionally or structurally important if it is evolutionarily conserved (Kimura, 1983). Therefore, change of the evolutionary conservation at a particular residue may indicate the involvement of functional divergence during the evolution of a gene family (Gu, 1999). Furthermore, Gu (2001) made a distinction between Type-I and Type-II functional divergences. Note that these two types of functional divergence may have other names. For instance, the basic Evolutionary Trace approach (Lichtarge et al. 1996; Madabushi et al. 2004) mainly focused on cluster-specific residues related to Type-II functional divergence. Gribaldo et al. (2003) also looked at Type-II functional divergence as called ‘constant-but-different’. Meanwhile, the weighted Evolutionary Trace approach proposed by Landgraf et al. (2001) was similar to Type-I functional divergence (Gu 1999).

Type-I functional divergence (Site-specific rate shift)

This type of functional divergence refers to the evolutionary process, resulting in site-specific rate shifts after gene duplication (Gu 1999; Gaucher et al. 2002b; Landgraf et al. 2001; Knudsen and Miyamoto 2001; Lopez et al. 1999). Typically, an amino acid residue is highly conserved in one duplicate gene, but highly variable in the other one. Gu

(1999) has developed a statistical method to test the significance of Type-I functional divergence between duplicate genes. Briefly, the two-state model proposed by Gu (1999) assumed that an amino acid residue (site) is in either one of two states: *related to functional divergence* if its evolutionary rate is shifted (up or down) after gene duplication; or *unrelated to functional divergence* for otherwise. The coefficient of (Type-I) functional divergence between duplicate genes, denoted by θ_I, is defined as the probability of being related to functional divergence. Clearly, a large value of θ_I indicates a high level of Type-I functional divergence, and *vice versa*. In a typical case when two gene clusters are generated by a gene duplication event, the coefficient of (Type-I) functional divergence between them can be estimated (Gu 1999; Gu and Vander Velden 2002). Rejection of the null hypothesis $\theta_I = 0$ means that the evolutionary rate has become different between the duplicate genes at some sites. Using this method, many case studies has demonstrated the functional-structural basis, e.g., the Caspase family (Wang and Gu 2001), and the Jak protein kinase family (Gu et al. 2002). Moreover, a site-specific profile based on the empirical posterior analysis is useful to predict amino acid residues that are crucial for functional divergence.

Type-II functional divergence (site-specific property shift)

As opposed to site-specific shift of evolutionary rate (Type-I functional divergence), Type-II functional divergence results in site-specific property shift. A typical case is that at a homologous residue (one column in the multiple alignment of the gene family), a radical shift of amino acid property, e.g., positively versus negatively charged, has occurred between two duplicate genes; otherwise they are both evolutionally conserved

within each of orthologous genes. Gu (2006) proposed a statistical method for the Type-II functional divergence inference by extending the two-state model (Gu 1999, 2001) to Type-II (cluster-specific) functional divergence: (*i*) In the early (*E*) stage after gene duplication, an amino acid residue can be in either of two states: Type-II unrelated and Type-II related. The probability of a residue being under the Type-II related status is denoted by θ_{II}, as called the coefficient of Type-II functional divergence. (*ii*) In the late (*L*) stage, any amino acid residue has no further Type-II functional divergence, so amino acid substitutions in this stage are mainly under purifying selection. Under the functional divergence unrelated status, the substitution model largely reflects the conserved evolution of protein sequences, which can be empirically determined by the Dayhoff model (Dayhoff et al. 1978), or the JTT model (Jones, Taylor, and Thornton 1992). In contrast, under the functional divergence-related status, radical amino acid substitutions may occur more frequently (Lichtarge et al. 1996). To avoid over-parameterization, Gu (2006) proposed a simple model that can distinguish between the radical and conserved amino acid substitutions. First, we tentatively classify twenty amino acids into four groups: charge positive (*K, R, H*), charge negative (*D, E*), hydrophilic (*S, T, N, Q, C, G, P*), and hydrophobic (*A, I, L, M, F, W, V, Y*). An amino acid substitution is called radical if it changes from one group to another; otherwise it is called conserved. In a typical case when two gene clusters generated by a gene duplication event, the coefficient of (Type-II) functional divergence between them can be estimated. Moreover, a site-specific profile based on the empirical posterior analysis is useful to predict amino acid residues that are crucial for Type-II functional divergence.

1.2.2 Evidence for the association between functional divergence and changes in function after gene duplication

Using statistical methods in software DIVERGE (Gu and Vander 2002), our group has conducted a large-scale analysis showing that site-specific rate shift (type-I functional divergence) is a general evolutionary pattern. Moreover, we found evidence that the level of site-specific rate shift of member genes could be related to protein structure differences (Wang and Gu, 2001; Gaucher et al. 2002a), the severity of knockout phenotypes, and tissue-specificity.

1.3 G-protein alpha subunit as an example

G proteins, short for guanine nucleotide binding proteins, are a family of proteins involved in second messenger cascades. These proteins are activated by G protein-coupled receptors and are made up of alpha, beta and gamma subunits. There are over 16 G-protein alpha subunits in animals, which can be further divided into four major classes: Gs, Gio, Gq, and G12, respectively (Simon et al. 1991; Neer 1995; Downes et al 1999; Cabrera-Vera et al. 2003), depending on their actions upon the effectors. In the project, we have predicted amino acid residues that are related to either Type-I or Type-II functional divergence between the Gq and Gs subfamilies. The inferred ancestral sequences for these sites are helpful to explore the trends of functional divergence. Finally, these predicted residues are mapped to the protein structures to test whether these residues may have 3D structure or solvent accessibility preference. We shall demonstrate

how to identify amino acid residues that are crucial for different types of functional divergence between duplicate genes, infer the trend of evolutionary changes at these residues, as well as protein structural interpretations of these predicted residues.

CHAPTER II. DATA AND METHODS

2.1 Animal G-protein alpha subunits family data set preparation

Protein sequence data

The starting set of vertebrate G-protein alpha subunits sequences were downloaded from the Homologous Vertebrate Genes Database (http://pbil.univ-lyon1.fr/databases/hovergen.html). A standard BLAST search against NCBI non-redundant protein sequence database added more vertebrate sequences and the invertebrate out group to the data set. Finally we got 81 amino acid sequences of animal G-protein alpha subunits. The multiple alignments were made using the program ClustalX. The phylogenetic tree of the whole family was inferred by the neighbor-joining (NJ) method (Saitou and Nei 1987). The parsimony (PAUP4.0) and likelihood (PHYLIP) methods give virtually the same topology.

Protein structure data

We downloaded the 3D structures of Gs and Gq from the RCSB Protein database (http://www.rcsb.org/), and the MMDB (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Structure) with respective data formats. In our study, we utilized the structure of Gs_BOVIN, the chain C of the PDB entry 1AZS, and Gq_MOUSE, the chain Q of PDB entry 2BCJ.

2.2 DIVERGE2: Analytical pipeline for comprehensive functional divergence analysis

The software DIVERGE (Gu and Vander Velden 2002) is a software system to study functional divergence between member genes of a protein family based on (site-specific) shifted evolutionary rates (Type-I) after gene duplication.

Posterior analysis results in a site-specific profile for predicting important amino acid residues that are responsible for this type of functional divergence. Moreover, when the 3D protein structure is available, these predicted amino acid residues can be mapped to the 3D structure viewer to explore their structural basis.

The updated version, named DIVERGE2, has provided more options (e.g., Type-II functional divergence) to explore functional evolution of protein family sequences. One can use the site-specific profiles to detect amino acid residues that are crucial for this type (I or II) of functional divergence. In practice, one may use the site (k)-specific score $Q_I(k)$, or $Q_{II}(k)$, the posterior probability that site k is related to Type-I or Type-II functional divergence. Another commonly used measure is based on the posterior ratio; in our case, it is given by $R_I(k) = Q_I(k)/[1-Q_I(k)]$, or $R_{II}(k) = Q_{II}(k)/[1- Q_{II}(k)]$. When a cutoff is given, important residues for two types of functional divergence are predicted.

2.3 Ancestral sequence inference

Ancestral sequence inference under a given phylogeny is becoming an important approach in molecular biology and functional comparative genomics (Golding and Dean,

1998). This is partly because evolution has selected proteins for function over hundred or even thousand millions of years, keeping those that carried out critical functions, and eliminating deleterious mutations. We have recognized that ancestral sequence reconstruction is a powerful technique for linking sequence to function. An important development of DIVERGE2 is to provide an analytical pipeline for combining functional divergence and ancestral sequence inference, which can be used to infer the trends of functional divergence. Currently, DIVERGE2 adopts the Bayesian algorithm of Zhang et al. (1997) to infer the ancestral sequences under a known phylogeny of gene family. It is a simplified version of Yang et al. (1995) in which the branch lengths of the phylogenetic tree are estimated using a least squared method rather than the maximum-likelihood method. Each site in the inferred ancestral sequence receives the assignment of amino acid with the highest posterior probability. Using this approach one may determine whether an amino acid residue that was highly conserved in the ancestral protein sequence now becomes highly variable, or *vice versa*.

2.4 Mapping predicted amino acid residues to protein structure

Important features of the physical environment of a residue such as secondary structure, and whether the site is on the surface or in the interior of a protein, can be extracted from the solved 3D structure. We take the criteria of accessible surface area (ASA) to define the surface area of a bimolecule accessible to a solvent (Lee and Richards 1971). Residues are considered to be solvent exposed (on the surface, **o** for short) or be buried (inside, **i** for short) according to the relative ASA in the protein. The program JOY

(Mizuguchi et al. 1998) was used to compute the relative ASA and assign all the residues to “solvent inaccessible” or “solvent accessible” with the default cutoff of 7% relative accessibility. As an alternative method for computing residue accessibility, NACCESS (Hubbard and Thornton 1993) gave similar results.

CHAPTER III. RESULTS AND DISSCUSSION

Consistent to previous results, the four major classes of G-protein alpha subunit family are monophyletic. As an example, we choose Gs and Gq, the two major classes of G-protein alpha subunits, to demonstrate the ancestral-based analysis of functional divergence. Fig.1 is the phylogenetic tree of Gs and Gq classes. The Gs class, which consists of Gs and Golf subtypes, is involved in hormonal stimulation of adenylate cyclase and opening of Ca2+ channels. While the Gs subtype is expressed in almost all tissue types, the Golf subtype is expressed exclusively in olfactory cells and is thought to be involved specifically in odorant signal transduction (Kaziro et al. 1991). On the other hand, the class Gq has the function of simulating phospholipase C (PLC), which has four subtypes: Gq, G11, G14, and G15. In spite of the fact that Gq and G11 are widely distributed and often found in the same cell types, they may have different receptors and effectors or act in different developmental stages. G14 and G15 are tissue-specific, which may interact with different members of the phospholipase family. In this demonstration, G15 subtype is not included in the Gq class. It should be noted that the alignment of all the 81 protein sequences is the input of our software; whereas Gs and Gq(without G15) classes are two demonstration clusters selected in the analysis.

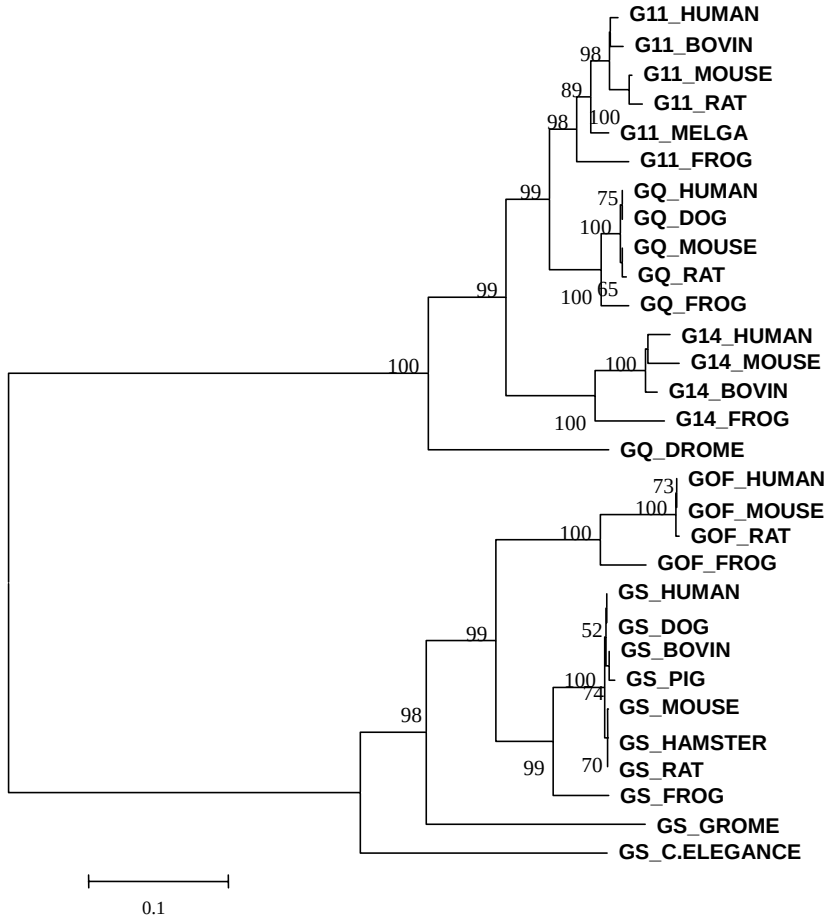

Fig. 1. The NJ tree of Gq and Gs based on the multiple sequence alignment with Poisson distance. Bootstrap score >50% are presented. The accession numbers for protein sequences are P50148(GQ_HUMAN), Q28294(GQ_DOG), P21279(GQ_MOUSE), P82471(GQ_RAT), P38410(GQ_FROG), P29992(G11_HUMAN), P38409(G11_BOVIN), P21278(G11_MOUSE), Q9JID2(G11_RAT), P45645(G11_MELGA), P43444(G11_FROG), O95837(G14_HUMAN), P38408(G14_BOVIN), P30677(G14_MOUSE), O73819(G14_FROG), JN0115(GQ_DROME), P04895(GS_HUMAN), CAA78161(GS_DOG), P04894(GS_MOUSE), AAA40827(GS_RAT), CAA35516GS_HAMSTER), P04896(GS_BOVIN), P29797(GS_PIG), CAA39571(GS_FROG), Q8CGK7(GOF_MOUSE), P38406(GOF_RAT), P38405(GOF_HUMAN), CAC82735(GOF_FROG), NP_477506(GS_DROME), NP_490817(GS_C.ELEGANCE).

3.1 Functional divergence between Gs and Gq proteins

Type-I functional divergence

We first tested the site-specific shift of evolutionary rate (Type-I functional divergence) after the gene duplication event leading to Gs and Gq subtypes. The coefficient of Type-I functional divergence between Gs and Gq is $\theta_I = 0.53 \pm 0.08$, which is significantly larger than 0. Hence, site-specific rate difference may occur at some amino acid residues after the gene duplication. Fig.2 shows the site-specific profile of posterior ratio, $R_I(k)$; notably, most sites are unlikely to be involved in the Type-I functional divergence. We used the cutoff $R_I > 2$ (the posterior probability $Q_I(k) > 0.67$) to identify the (Type-I) functional divergence-related residues between Gs and Gq, and obtained twenty-five amino acid residues (Fig.3). This cutoff value is empirical. Generally, the cutoff value in terms of posterior ratio is large than 1(large than 0.5 for *Q*, the posterior probability); for large θ values, we should choose a large cutoff value to avoid too much false positive results. These sites clearly show a typical pattern of Type-I functional divergence, i.e., conserved amino acid in one cluster, and diverse amino acids in the other one. Moreover, these predicted sites can be divided into two groups. Group A in Fig. 3 includes 15 sites that conserved in Gq but not conserved in Gs, while the group B includes 10 sites that conserved in Gs but not conserved in Gq. Consequently, Gq proteins become more conserved than Gs proteins.

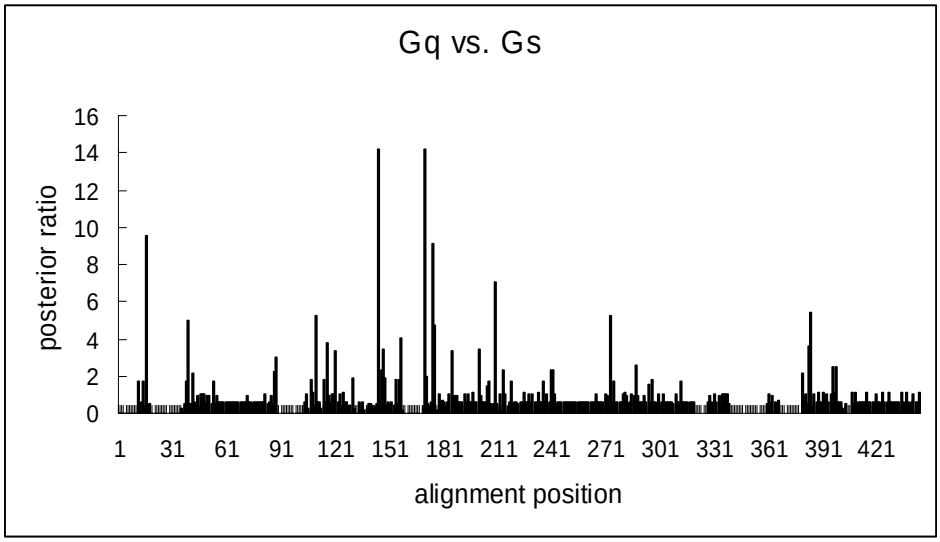

Fig. 2. The site-specific profile for predicting critical amino acid residues responsible for the Type-I functional divergence between clusters Gq and Gs measured by posterior ratio $R_I(k)$.

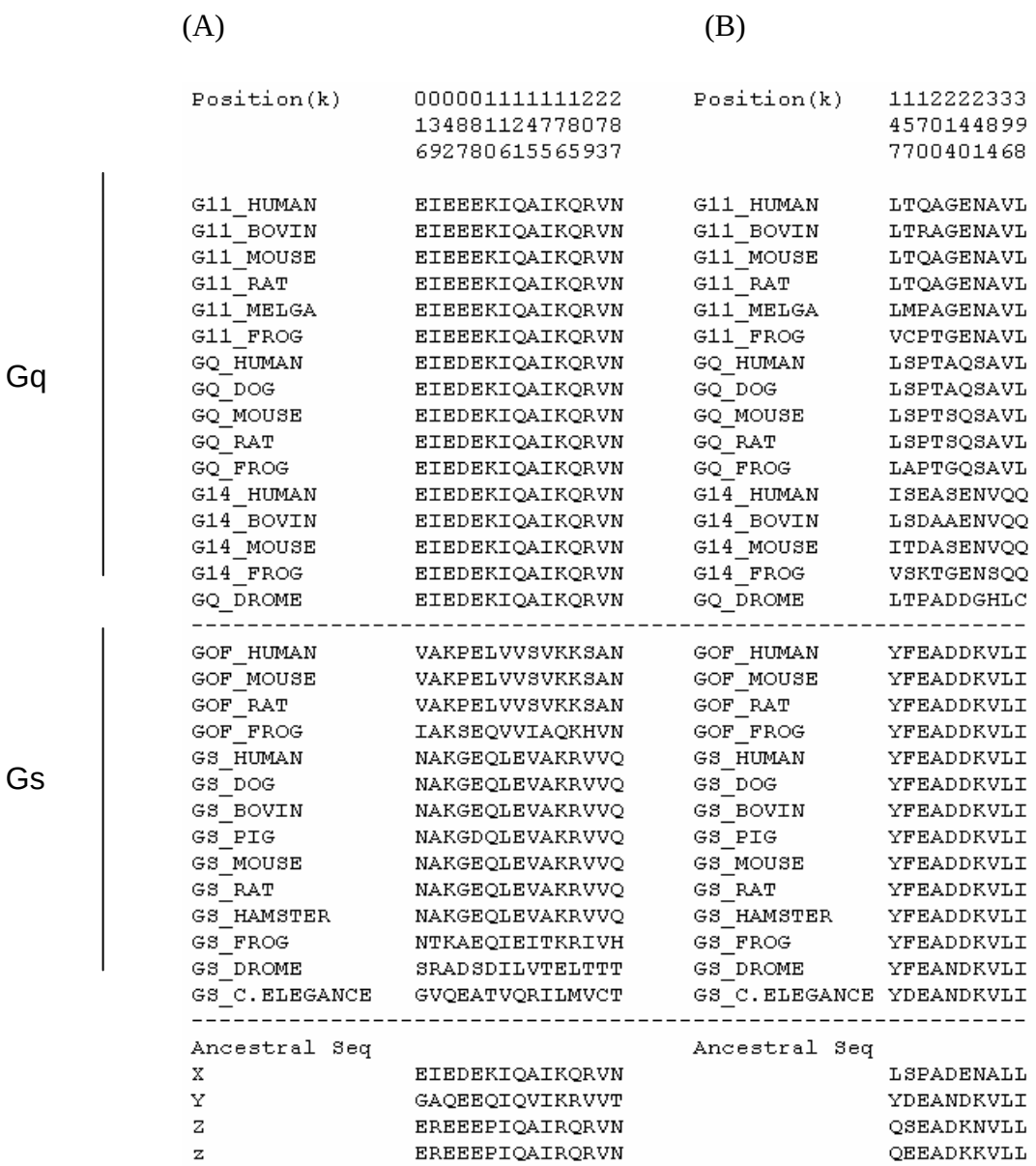

Fig. 3. Type-I functional divergence related amino acid sites candidates and ancestral sequence inference of these sites. (A): Category I: amino acids conserved in Gq cluster but variable in Gs cluster. (B): Category II: amino acids conserved in Gs cluster but variable in Gq cluster. X: The ancestral amino acids for the Gq cluster at the candidate sites. Y: The ancestral amino acids for the Gs cluster at the candidate sites. Z: The ancestral amino acids for the common ancestor of Gs and Gio cluster at the candidate sites. z: The ancestral amino acids for the common ancestor of the Gq and G12 cluster at the candidate sites.

Type-II functional divergence

Based on the same multiple alignment of protein sequences, we obtained the estimate of the coefficient of Type-II functional divergence, θ_{II} = 0.325±0.055, between the Gs and Gq alpha proteins, which is significantly larger than 0. It suggests that, after the gene duplication, some amino acid residues that are evolutionarily conserved in both Gs and Gq proteins may have radical changes in their amino acid properties. Fig.4 shows the site-specific profile based on the posterior ratio, $R_{II}(k)$, for Type-II functional divergence between Gs and Gq proteins. Notably, most residues receive very low scores, indicating that only a small portion of amino acid residues that have involved in this type of functional divergence. 29 amino acid residues with the highest scores (the posterior ratio R_{II} >17) show a typical shift of amino acid properties at conserved residues (Fig.5), as been demonstrated in Table 1.

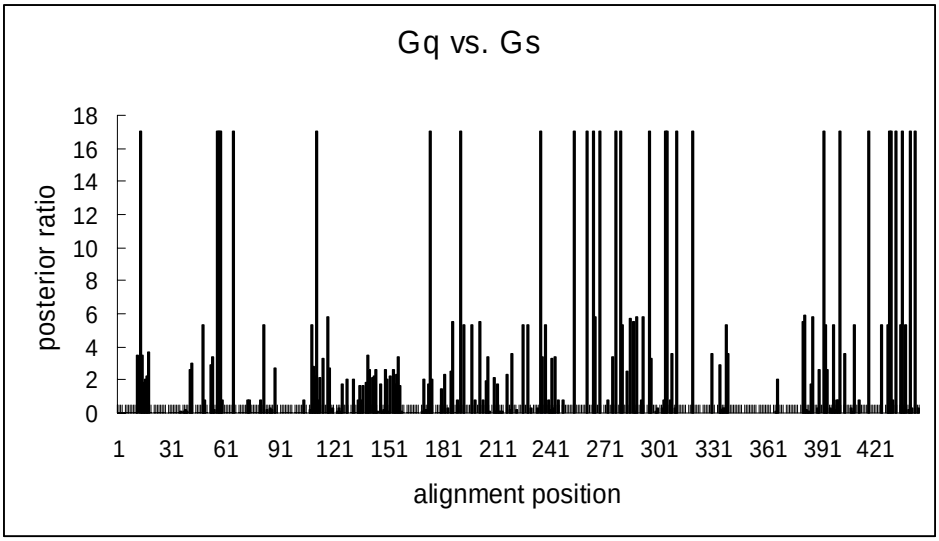

Fig. 4. The site-specific profile for predicting critical amino acid residues responsible for the Type-II functional divergence between clusters Gq and Gs measured by posterior ratio $R_{II}(k)$.

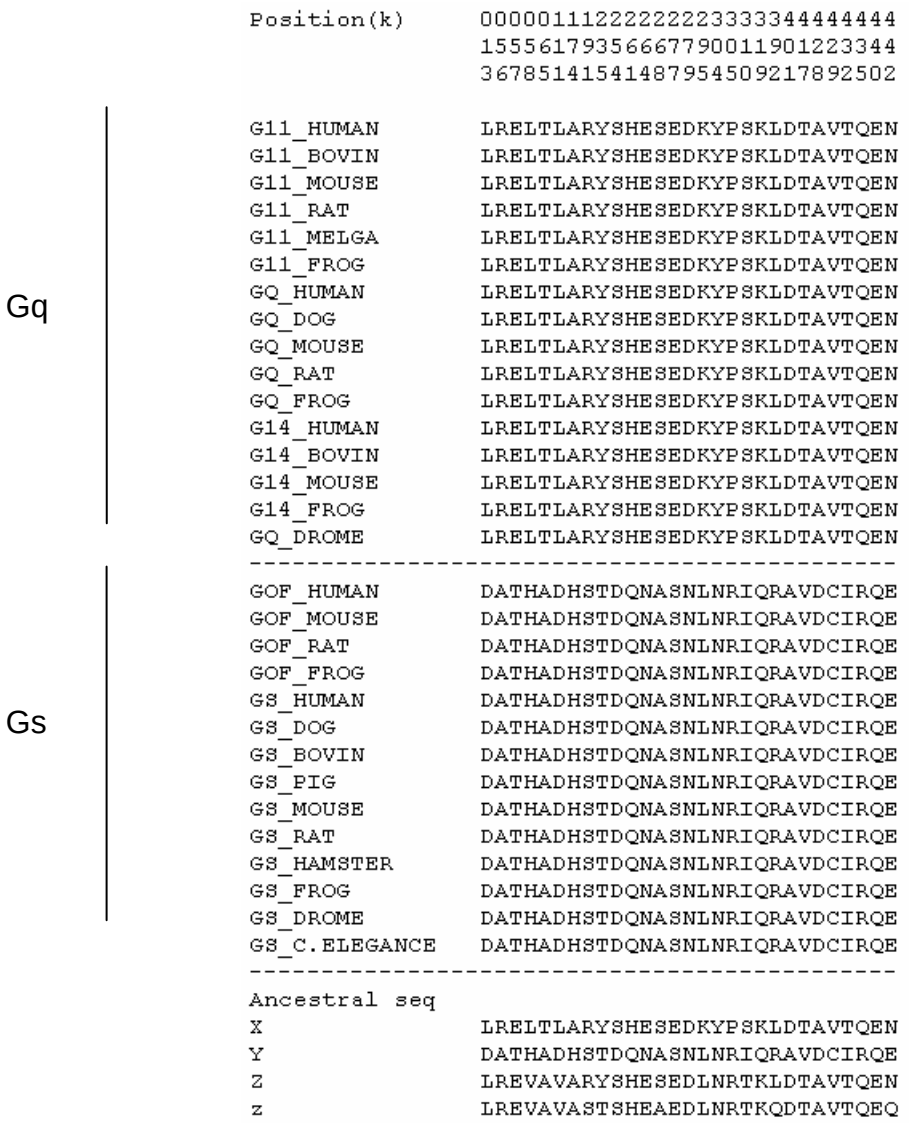

Fig. 5. Type-II functional divergence related amino acid sites candidates and ancestral sequence inference of these sites. X, Y, Z, and z are the same definitions as in the Fig. 3.

Table 1. Overview of the amino acid changes in the predicted 29 sites in Type-II functional divergence.

alignment Position	Gq	Gs	Propertie change
13	L	D	hydrophobic/-
56	R	A	+/ hydrophobic
57	E	T	-/ hydrophilic
58	L	H	hydrophobic/+
65	T	A	hydrophilic/ hydrophobic
111	L	D	hydrophobic/-
174	A	H	hydrophobic/+
191	R	S	+/ hydrophilic
235	Y	T	hydrophobic/ hydrophilic
254	S	D	hydrophilic/-
261	H	Q	+/ hydrophilic
264	E	N	-/ hydrophilic
268	S	A	hydrophilic/ hydrophobic
277	E	S	-/ hydrophilic
279	D	N	-/ hydrophilic
295	K	L	+/ hydrophobic
304	Y	N	hydrophobic/ hydrophilic
305	P	R	hydrophilic/+
310	S	I	hydrophilic/ hydrophobic
319	K	Q	+/ hydrophilic
392	L	R	hydrophobic/+
401	D	A	-/ hydrophobic
417	T	V	hydrophilic/ hydrophobic
428	A	D	hydrophobic/-
429	V	C	hydrophobic/ hydrophilic
432	T	I	hydrophilic/ hydrophobic
435	Q	R	hydrophilic/+
440	E	Q	-/ hydrophilic
442	N	E	hydrophilic/-

3.2 Evolutionary trends of functional divergence - Ancestral inference analysis

Using the Bayesian ancestral sequence inference implemented in the software DIVERGE2, we inferred the ancestral sequences of all internal nodes on the phylogeny of G-protein alpha subunits, which provides further information about the evolutionary trends of functional divergence. Since our study is focused on the Gq and Gs clusters, we are interested in the ancestor node for the Gs cluster, Gq cluster, the Gs and Gio clusters, and the Gq and G12 clusters. All these four major internal nodes are represented in Fig. 6, where the “X” stands for the ancestor for the Gq cluster, the “Y” stands for the ancestor for the Gs cluster, the “Z” stands for the common ancestor for the Gs and Gio clusters, and the “z” stands for the common ancestor for the Gq and G12 clusters. The whole sequences for these four ancestors are concerned. In particular, the inferred ancestral amino acid residues related to Type-I or Type-II functional divergence sites are presented in Fig. 3 and Fig. 5, respectively, where X is the ancestral residues for the Gq cluster, Y is the ancestral residues for the Gs cluster, Z is the common ancestral residues for the Gs and Gio clusters, and z is the common ancestral residues for the Gq and G12 clusters.

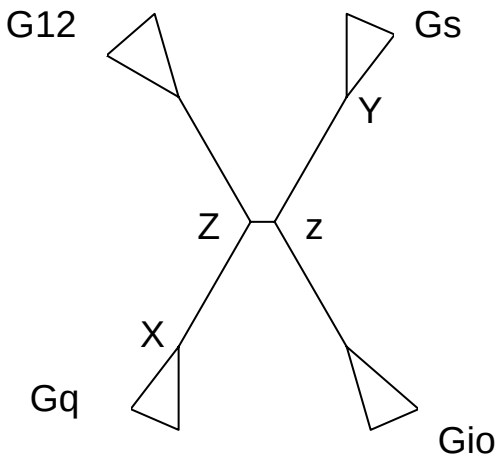

Fig. 6. The ancestral inference points for the G-protein alpha subunit family.

X: The ancestor for the Gq cluster. Y: The ancestor for the Gs cluster. Z: The common ancestor of Gs and Gio clusters. z: The common ancestor of the Gq and G12 clusters.

As shown in Fig. 3, there are two groups of Type-I functional divergence. Among fifteen residues in Group A, i.e., conserved in Gq but variable in Gs, five residues (e.g., position 088) show the conserved Gq-type amino acid at all four major internal nodes (X, Y, Z, and z), while six residues (e.g., position 016) show the conserved Gq-type amino acid in three internal nodes but not at the Y (common ancestor of Gs). Putting these two ancestral patterns together, it appears that the ancestral states of these eleven residues are all conserved Gq-type, whereas the variable Gs-type residues are the derived characters that are only specific to Gs-proteins. The only difference between these two ancestral patterns is that Type-I functional divergence at the first five residues likely occurred after the common ancestor of Gs (Y), while those at the second six residues occurred before the

node Y. On the other hand, among ten residues in Group B, i.e., conserved in Gs but variable in Gq, two residues (positions 200 and 396) show the conserved Gs-type amino acid at all four major internal nodes, while another two residues (positions 170 and 384) show such pattern except for the common ancestor of Gq (X). Interestingly, for the two residues at positions 214 and 398, the conserved Gs-type is recently derived, which is specific to the Gs cluster. For the rest of residues, one can not determine the trend of functional divergence, due to the statistical uncertainty of phylogenetic inference or ancestral sequence inference.

In the same manner, we examined the ancestral amino acid residues for Type-II functional divergence (Fig.5). Among twenty-nine predicted Type-II divergence related residues, seventeen residues (e.g., position 013) show an ancestral patterns indicating these amino acid property-shifts at conserved residues may occur in the evolutionary trend that can be simply represented as from the internal nodes z (ancestral type) to Y (Gs-type). In contrast, the ancestral pattern of four residues (e.g., position 065) indicates the evolutionary trend of Type-II functional divergence from the internal nodes Z (ancestral type) to X (Gq-type).

3.3 Protein structure mapping

Comparative study of molecular sequences and protein structures has provided many insights into protein folding, stability and evolution (Golding and Dean, 1998). The structure information for particular residues such as functional divergence related residues

could provide a deep insight into the evolution trends. The crystal structures of G-proteins, Gs and Gq, are both determined (Tesmer et al. 1997; Tesmer et al. 2005; Wall et al. 1995; Lambright et al. 1996; Cabrera-Vera et al. 2003), which provide the structural basis to investigate the functional interpretations of these Type-I and Type-II predicted residues. The software DIVERGE2 we have developed is capable of mapping a subset of amino acid residues onto the protein structure. It should be noted that some residues in either N or C termini are not available or simply disordered in the solved protein structures (Cabrera-Vera et al. 2003). Consequently, predicted Type-I and Type-II residues (sites) in these regions have to be excluded in the sequence-structure study.

A G-protein alpha subunit contains two domains: a GTPase domain involved in the binding and hydrolysis of GTP, and a helical domain connected to the GTPase domain by two linker regions (Cabrera-Vera et al. 2003). Fig. 7 shows the location of predicted Type-I and Type-II functional divergence related sites on the protein 3D structures of Gs_BOVIN and Gq_MOUSE. The mapping result suggests that the Type-I and Type-II sites have different patterns of location distribution. Among twenty-five Type-I sites predicted at the cutoff posterior ratio R_I>2, there are thirteen sites in the helical domain (in the range of PDB sites 86 - 202 in Gs, or PDB sites 70-185 in Gq), while the rest, around half of the total, are located in the GTPase domain. It should be noted that some predicted sites in N-terminal can not be mapped due to lacking of structure information as mentioned in previous paragraph, therefore Fig. 7 only shows part of Type-I sites in the GTPase domain. Interestingly, in the twenty-nine Type-II (radical cluster-specific) sites

predicted at the posterior ratio cutoff 17 (or posterior probability 0.94), there are only three sites located in the helical domain, while twenty-six sites are in the GTPase domain (p<0.001, binomial test). Fig.7 shows the locations of twenty-eight sites out of the predicted twenty-nine Type-II sites and reveals that most of the predicted Type-II sites are located in the ATPase domain. Cabrera-Vera et al. (2003) mentioned that the helical domain is the most divergent domain among the G-protein alpha subunits, whereas the GTPase domain is much more conserved. Our result indicates that the Type-II sites may be involved in more conserved domain, implying the relevance of these sites to the domain function.

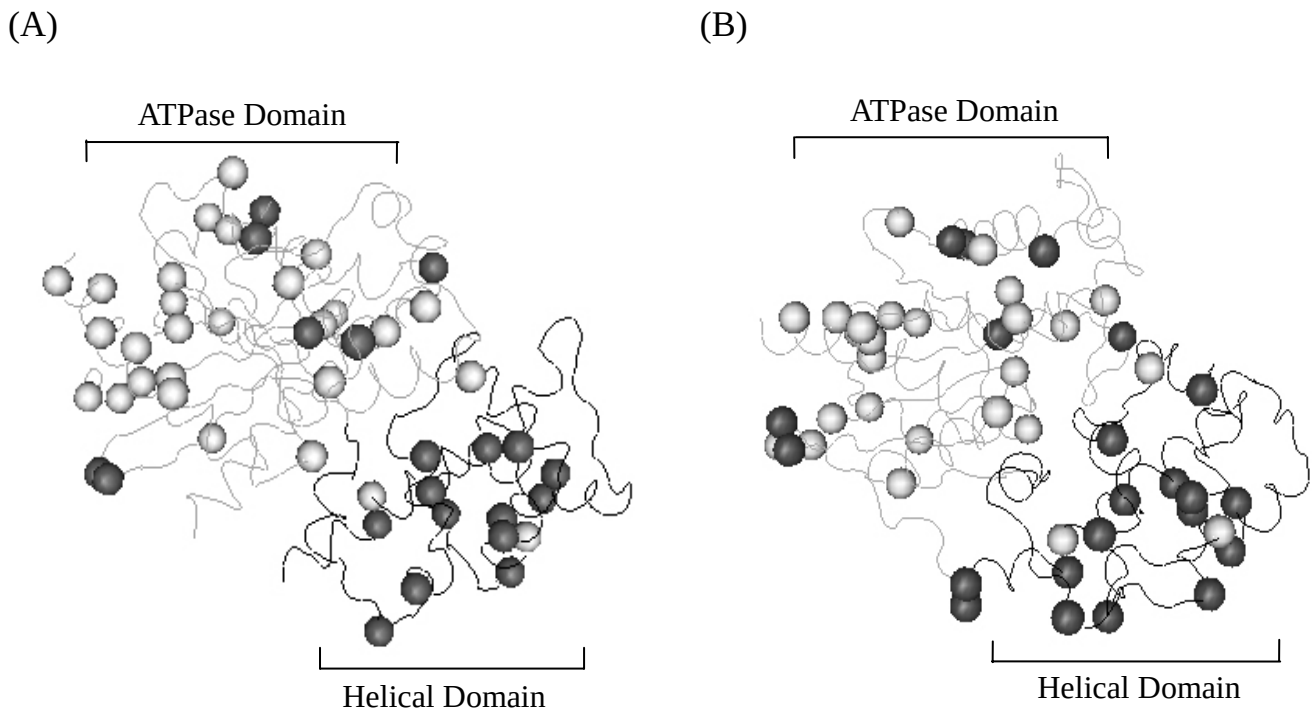

Fig. 7. The mapping of the predicted Type-I and Type-II sites onto protein 3D structures

(A): The mapping of the predicted Type-I and Type-II sites onto the Gs-BOVIN protein structure (Chain C in the structure 1AZS). (B): The mapping of Type-I and Type-II sites onto the Gq-MOUSE structure (Chain Q in the structure 2BCJ). The dark ball represents the predicted Type-I functional divergence-related residues; the light ball represents the predicted Type-II functional divergence-related residues. The gray domain is ATPase domain of the G protein alpha subunit; the black domain is helical domain of the G-protein alpha subunit.

Moreover, we examined the distribution of Type-I and Type-II functional divergence related sites in the secondary structure elements. Table 2 (A, and B) shows that most of these sites are within the typical repeat elements of secondary structure such as alpha helices, beta strands and 310 helices.

Besides the secondary structure distribution of these functional divergence-related sites, we are concerned about whether they are located on the surface or in the interior of the protein. To address this issue, we take the criteria of accessible surface area (ASA) to define the surface area of a bimolecule accessible to a solvent (Lee and Richards 1971). Residues are considered to be solvent exposed (on the surface, **o** for short) or be buried (inside, **i** for short) according to the relative ASA in the protein. Overall, we found that the Type-I and the Type-II sites may have the tendency to be located in the surface area (Table 2 C), though the differences are not statistically significant (p-value >0.05). Moreover, the accessibility of most Type-I sites, except for three sites(positions 121, 145 and 157), has no changes between Gq and Gs, while in the Type-II sites, eight sites are found to have accessibility changes between the Gq and Gs proteins. Interestingly, there are more solvent accessible residues in Gs than in Gq among these Type-II sites.

Table 2. The structure features of the Type-I and Type-II functional divergence related residues in Gq and Gs: the secondary structure conformation and the solvent accessibility.

A: Type-I functional divergence related sites and their structure features.

Alignment position	Residue		Secondary structure		Solvent accessibility	
	Gq	Gs	Gq	Gs	Gq	Gs
16	e	n	-	-	-	-
39	i	a	-	-	-	-
42	e	k	-	-	-	-
87	d	g	a	-	o	-
88	e	e	a	-	o	-
110	k	q	a	a	o	o
116	i	l	a	a	i	i
121	q	e	a	a	i	o
145	a	v	a	a	i	o
175	i	a	a	a	i	i
176	k	k	a	a	o	o
185	q	r	a	a	o	o
209	r	v	a	a	o	o
273	v	v	b	b	i	i
287	n	q			o	o
147	l	y	a	a	o	o
157	s	f			o	i
170	p	e	a	a	o	o
200	t	a	a	a	i	i
214	s	d			o	o
240	q	d			o	o
241	s	k			o	o
384	a	v	a	a	o	o
396	v	l	a	a	o	o
398	l	i	a	a	o	o

B: Type-II functional divergence related sites and their structure features.

Alignment position	Residue		Second structure		Solvent accessibility	
	Gq	Gs	Gq	Gs	Gq	Gs
13	L	D	-	-	-	-
56	R	A			o	o
57	E	T	b	b	o	o
58	L	H	b	b	o	i
65	T	A			i	o
111	L	D	a	a	o	o
174	A	H	a	a	i	o
191	R	S		3	o	o
235	Y	T	b	b	o	o
254	S	D		3	o	o
261	H	Q	3	3	o	o
264	E	N			o	o
268	S	A	b	b	i	i
277	E	S	3	3	i	i
279	D	N	3	3	i	o
295	K	L	a	a	o	o
304	Y	N			i	o
305	P	R	3		i	o
310	S	I			i	o
319	K	Q	a	a	o	o
392	L	R	a	a	o	o
401	D	A			o	o
417	T	V			i	i
428	A	D	a	a	o	o
429	V	C	a	a	i	i
432	T	I	a	a	o	i
435	Q	R	a	a	o	o
440	E	Q	-	a	-	o
442	N	E	-	a	-	o

In this table, the letter “a” stands for alpha helices, “b” stands for beta strands and “3” stands for 310 helices.

“i ”means solvent inaccessible and “o” means solvent accessible.

“- ” means there is no residue structure information available.

C: the comparison of the solvent accessibility of Type-I, Type-II functional divergence related residues to all the residues in the sequence.

	Gs			Gq		
	TypeI	TypeII	Whole sequence	TypeI	TypeII	Whole sequence
#sites	20	28	339	22	26	317
#Accessible sites	15	22	238	16	16	198
Accessible /total	0.75	0.79	0.70	0.73	0.62	0.62

The table only considered the residues with available structure information, i.e., some N and C termini are not included. The p-value for testing the non-difference of the accessibility between the functional divergence related (Type-I or Type-II) sites and the functional divergence unrelated sites is large than 0.1 and reveals no significant difference between those sites.

CHAPTER IV. CONCLUSIONS

Providing substantial genomic data, plus powerful computational tools, it is now desirable to develop a comprehensive analytical pipeline to perform functional divergence analysis of protein families. In this project, we use animal G-protein alpha subunit family as an example to illustrate such an analytical pipeline. Advanced to our previous works, which can only detect type-I functional divergence between duplication genes (or subfamilies) in protein families and identify residues responsible for the functional divergence, this approach includes detecting different types, say, type-I and type-II, of functional divergence, as well as identifying the functional residues, and can be further combined with ancestral sequence inference and available residue protein structure information.

By this approach, we are more than able to exam two types of functional divergences between duplication genes. With the combination of functional divergence analysis and the ancestral sequence inference, we are able to trace the evolutionary trend of two types of functional divergence of amino acid residues after the gene duplication. With the sequence-3D structure mapping we can get the structure features of the particular functional divergence of amino acid residues, and explore the sequence-structure relationship during the evolution. Clearly, these pieces of evolutionary information are useful for making testable hypothesis about functional divergence between subtypes of G-protein alpha subunits, which can be verified by further experimentation.

REFERENCES CITED

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28(1):235-242.

Bork P, Koonin EV. 1998. Predicting functions from protein sequences--where are the bottlenecks? Nat Genet 18(4):313-318.

Cabrera-Vera TM, Vanhauwe J, Thomas TO, Medkova M, Preininger A, Mazzoni MR, Hamm HE. 2003. Insights into G protein structure, function, and regulation. Endocr Rev 24(6):765-781.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff, M. O. (ed.) Atlas of Protein Sequence and Structure, vol. 5, (suppl. 3),345-352. National Biomedical Research Foundation, Washington DC.

Doolittle RF. 1995. The origins and evolution of eukaryotic proteins. Philos Trans R Soc Lond B Biol Sci 349(1329):235-240.

Downes GB, Gautam N. 1999. The G protein subunit gene families. Genomics 62(3):544-552.

Eisen JA, Fraser CM. 2003. Phylogenomics: intersection of evolution and genomics. Science 300(5626):1706-1707.

Gao X, Vander Velden KA, Voytas DF, Gu X. 2005. SplitTester: software to identify domains responsible for functional divergence in protein family. BMC Bioinformatics 6:137.

Gaucher EA, Das UK, Miyamoto MM, Benner SA. 2002a. The crystal structure of eEF1A refines the functional predictions of an evolutionary analysis of rate changes among elongation factors. Mol Biol Evol 19(4):569-573.

Gaucher EA, Gu X, Miyamoto MM, Benner SA. 2002b. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci 27(6):315-321.

Golding GB, Dean AM. 1998. The structural basis of molecular adaptation. Mol Biol Evol 15(4):355-369.

Gribaldo S, Casane D, Lopez P, Philippe H. 2003. Functional divergence prediction from evolutionary analysis: a case study of vertebrate hemoglobin. Mol Biol Evol 20(11):1754-1759.

Gu J, Wang Y, Gu X. 2002. Evolutionary analysis for functional divergence of Jak protein kinase domains and tissue-specific genes. J Mol Evol 54(6):725-733.

Gu X. 1999. Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16(12):1664-1674.

Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. Mol Biol Evol 18(4):453-464.

Gu X. 2006. A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. Mol Biol Evol 23(10):1937-1945.

Gu X, Vander Velden K. 2002. DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics 18(3):500-501.

Henikoff S, Greene EA, Pietrokovski S, Bork P, Attwood TK, Hood L. 1997. Gene families: the taxonomy of protein paralogs and chimeras. Science 278(5338):609-614.

Hubbard, SJ, Thornton, JM. 1993. 'NACCESS', Computer Program, Department of Biochemistry and Molecular Biology, University College London.

Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. Nature 358(6381):86-89.

Jordan IK, Bishop GR, Gonzalez DS. 2001. Sequence and structural aspects of functional diversification in class I alpha-mannosidase evolution. Bioinformatics 17(10):965-976.

Kaziro Y, Itoh H, Kozasa T, Nakafuku M, Satoh T. 1991. Structure and function of signal-transducing GTP-binding proteins. Annu Rev Biochem 60:349-400.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Knudsen B, Miyamoto MM. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci U S A 98(25):14512-14517.

Lambright DG, Sondek J, Bohm A, Skiba NP, Hamm HE, Sigler PB. 1996. The 2.0 A crystal structure of a heterotrimeric G protein. Nature 379(6563):311-319.

Landgraf R, Xenarios I, Eisenberg D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol 307(5):1487-1502.

Lee B, Richards FM. 1971. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55(3):379-400.

Li WH, Gu Z, Wang H, Nekrutenko A. 2001. Evolutionary analyses of the human genome. Nature 409(6822):847-849.

Lichtarge O, Bourne HR, Cohen FE. 1996. An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257(2):342-358.

Lopez P, Forterre P, Philippe H. 1999. The root of the tree of life in the light of the covarion model. J Mol Evol 49(4):496-508.

Madabushi S, Gross AK, Philippi A, Meng EC, Wensel TG, Lichtarge O. 2004.

Evolutionary trace of G protein-coupled receptors reveals clusters of residues that determine global and class-specific functions. J Biol Chem 279(9):8126-8132.

Mizuguchi K, Deane CM, Blundell TL, Johnson MS, Overington JP. 1998. JOY: protein sequence-structure representation and analysis. Bioinformatics 14(7):617-623.

Neer EJ. 1995. Heterotrimeric G proteins: organizers of transmembrane signals. Cell 80(2):249-257.

Nei M. 1987. Molecular evolutionary genetics.Columbia University Press, New York.
Ohno S. 1970. Evolution by gene duplication Springer-Verlag, Berlin.

Rastogi S, Liberles DA. 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol 5(1):28.

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

Sidow A. 1996. Gen(om)e duplications in the evolution of early vertebrates. Curr Opin Genet Dev 6(6):715-722.

Simon MI, Strathmann MP, Gautam N. 1991. Diversity of G proteins in signal transduction. Science 252(5007):802-808.

Tesmer JJ, Sunahara RK, Gilman AG, Sprang SR. 1997. Crystal structure of the catalytic domains of adenylyl cyclase in a complex with Gsalpha.GTPgammaS. Science 278(5345):1907-1916.

Tesmer VM, Kawano T, Shankaranarayanan A, Kozasa T, Tesmer JJ. 2005. Snapshot of

activated G proteins at the membrane: the Galphaq-GRK2-Gbetagamma complex. Science 310(5754):1686-1690.

Wall MA, Coleman DE, Lee E, Iniguez-Lluhi JA, Posner BA, Gilman AG, Sprang SR. 1995. The structure of the G protein heterotrimer Gi alpha 1 beta 1 gamma 2. Cell 83(6):1047-1058.

Wang Y, Gu X. 2001. Functional divergence in the caspase gene family and altered functional constraints: statistical analysis and prediction. Genetics 158(3):1311-1320.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141(4):1641-1650.

Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol 44 Suppl 1:S139-146.

Zhou H, Gu J, Lamont SJ, Gu X. 2006. Evolutionary analysis for functional divergence of Toll-like receptor gene family and altered functional constraints. J Mol Evol (in press).

ACKNOWLEDGEMENTS

I would first like to thank my major professor, Dr. Xun Gu, for being my mentor, supporting me, offering me great guidance on my research projects, and always giving me encouragements during my graduate studies.

I would also like to thank my co-major professor, Dr. Karin Dorman, for kindly discussing with me and providing great suggestions on both my research projects and graduated studies. I must thank Dr. Xiaoqiu Huang, my committee professor, your guidance and help was very much appreciated. Thank you all for being on my committee.

Next, I would give special thanks to Dongping Xu for developing the DIVEGE2 software and giving me valuable suggestions on the usage of software and analysis of the data. Special thanks to Dr. Jianying Gu for introducing me into the research area of gene duplication functional divergence.

Then, I would like to thank all my former and current labmates, Dr. Zhongqi Zhang, Dr. Shiquan Wu, Dr. Zhixi Su and Yong Huang. I will forever enjoy our deep friendship and appreciate your help during my studies.

Last, I would like to thank my dear parents and husband. It is with their supports and encouragements that I could have the confidence on my abroad studies.