DOMAIN-BASED APPROACHES TO UNDERSTANDING PHYLOGENY AND ORTHOLOGY

BY

MAO-FENG GER

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biophysics and Computational Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2011

Urbana, Illinois

Doctoral Committee:

Professor Eric G. Jakobsson, Chair
Associate Professor Claudio F. Grosman
Assistant Professor Sheng Zhong
Assistant Professor Jian Ma

ABSTRACT

Domain-based approaches are used in phylogenetic reconstruction and functional identification. Two groups of ionotropic glutamate receptors (iGluR's) were identified with the topology of the binding core and pore-loop of the eukaryotic iGluR's. Group 1 has a potassium-like selectivity filter and Group 2 is most closely related to eukaryotic iGluR's. The relationship among them was investigated in this research. Then, the domain complexity of proteins was analysed on a comprehensive basis. Our results showed that bacterial and archaeal proteins are as complex as eukaryotic proteins in domain abundance, but more promiscuous. Proteins emerged in early stage are also more promiscuous, but with low domain abundance. The possible application of protein comparison based on domain content was also suggested in this research and could be used to help the identification of function and orthology. Therefore, domain-based approaches are proved to be useful in many areas of proteome research, including functional annotation, evolutionary illustration, and protein-protein network construction.

ACKNOWLEDGEMENTS

First, and foremost, I would like to express sincere appreciation for the advice, assistance, and patience shown throughout this research by my dear advisor, Dr. Eric G. Jakobsson. His capable and friendly counsel, unfailing interest, and continuing encouragement made this work possible.

A hearty appreciation is also extended to Dr. Claudio F. Grosman, Dr. Sheng Zhong, and Dr. Jian Ma for serving on my thesis committee and for providing helpful suggestions of this thesis.

There are many people I must thank in Jakobsson's group for providing their thoughts and suggestions in the process of this research. They extensively contributed to this thesis by invaluable and professional knowledge.

Special thanks to Cindy Dodds and the faculty members in Biophysics for helping me in many ways. Thanks also to my friends who constantly cheered me by offering their kind, lovely inspiration.

Finally, I would like to thank my family for their unconditional support encouragement, which helped me in more ways than they know as I pursued my goal. The deepest gratitude is to my parents, who devoted their time and love to make my education the highest priority.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION ----------1

Overall description of thesis ----------1

UniProt databases ----------2

Domain databases ----------3

Practical extraction and report language (Perl) ----------5

Biology WorkBench and several bioinformatics tools ----------6

Cytoscape ----------8

References ----------9

CHAPTER 2: DOMAIN-BASED IDENTIFICATION AND ANALYSIS OF GLUTAMATE RECEPTOR ION CHANNELS AND THEIR RELATIVES IN PROKARYOTES ----------11

Abstract ----------11

Introduction ----------12

Materials and methods ----------15

Results ----------17

Features and evolution of the prokaryotic glutamate receptor channels -----17

Sequence analysis of group 1 and group 2 sequences ----------21

Phylogenetic analysis of group 1 and group 2 sequences ----------21

Comparison with eukaryotic glutamate receptor channels ----------22

Discussions ----------23

References ----------25

Tables ----------29

Figures ----------34

CHAPTER 3: EVOLUTION OF DOMAIN COMPLEXITY IN PROTEINS ----------42

Abstract ----------42

Background ----------42

Results ----------46

Domain content retrieval ----------46

Domain-domain network and degree of connectivity ----------48

The analysis of domain compositions ----------50

The analysis of domain content in 24 eukaryotic proteomes ----------52

Implications of combination of domains ----------53

Discussions ----------55

Domain abundance ----------55

Domain connectivity ----------56

Capability to form versatile domain combinations ----------57

The evolutionary history of domain rearrangement ----------58

Methods ----------58

References ----------62

Tables ----------66

Figures ----------69

CHAPTER 4: THE COMPARISON OF PROTEINS BY DOMAIN CONTENT AND ITS APPLICATION ----------78

Abstract ----------78

Background ----------78

Results ----------81

Computation of inverse domain frequency scores, domain evolutionary significance scores for domains and cosine similarity scores for pairs of domain compositions ----------81

The relatedness between cosine similarity scores by InterPro and Orthologous Matrix (OMA) ----------83

A case study: a network of 16 ion channels ----------84

Discussions ----------85

Methods ----------87

References ----------90

Tables ----------93

Figures ----------97

CHAPTER 5: FUTURE WORK ----------99

The importance of building a model of domain evolution ----------99

Preliminary thoughts about the model ----------100

The application of the domain evolution model ----------101

References ----------101

CHAPTER 1:

INTRODUCTION

Overall description of thesis

The core idea of this thesis is that the level of organization comprised of protein functional domains is as significant as the finer level of amino acids and bases, and the coarser level of complete proteins and genes. This is becoming more recognized than previously, but there is still significantly more analysis of protein evolution at the amino acid and protein levels than at the domain level. Thus I have explored the possible uses of domains in phylogenetic reconstruction and functional identification, and have examined very large-scale trends in domain-level evolution in the three major superkingdoms of life.

The following sections in chapter 1 are background knowledge of some significant computational tools and databases that will be used in the following chapters.

UniProt databases

The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data (http://www.uniprot.org/). The UniProt Knowledgebase (UniProtKB), comprising two sections: Swiss-Prot (manually annotated and reviewed) and TrEMBL (automatically annotated and is not reviewed), is the central hub for the collection of functional information on proteins among its four components [1, 2]. UniProtKB manages the core data for each protein entry (the amino acid sequence, protein name or description, taxonomic data, citation information...etc.), as well as annotation information (biological ontologies, cross-references to other biological databases…etc.). It also minimizes the redundancy to improve data quality. There are over 13.5 million entries in UniProtKB as of release 2011_01 of 11 January 2011 with 524 420 entries in UniProtKB/Swiss-Prot and 13 069 501 entries in UniProtKB/TrEMBL.

The data integration of UniProtKB ensures that information related to a protein is captured in the most appropriate resource and cross-reference to other databases is provided as much as possible. Therefore, with its unified view of

protein sequence and functional information, it can be provided as a data resource for the research and of proteomes.

Domain databases

Domains in proteins are the basic units of function, structure, evolution. Domain definitions can be produces by different approaches, such as regular expressions, profiles or hidden Markov models [3, 4]. There are a number of public domain databases, each one with a different focus. To utilize them, InterPro domain database is developed to integrates protein signatures from CATHGene3D, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY, and TIGRFAMs into one integrated resource [5]. By this integration, it achieves greater sequence and taxonomic coverage than any member database. All signatures representing the equivalent domain from its member database are merged into single InterPro entries with annotation describing the domain. Each InterPro entry is manually curated and supplemented with additional cross-reference to other biological database, such as Gene Ontology (GO) annotation, taxonomy of matching proteins …etc.

Unlike other database, the InterPro domain entries sometimes are redundant (i.e. overlapping) for the same protein region and these related InterPro entries are built into a hierarchic relationship (parent/child and contains/found in). Also, InterPro reported domain definition does not display repetitive domains within the same protein, so that there is only one domain for each entry can be found within a protein. Nowadays, InterPro domain definition is used for automatic annotation of the UniProtKB/TrEMBL database.

In order to include different natures of different domain databases, two other domain definitions were explored in this research, Pfam [6] and Gene3D [7]. The Pfam domain definition was generated from multiple sequence alignment by hidden Markov models. PfamA models are high quality and manually curated, whereas PfamB models are automatically generated. The Gene3D domain definition was derived from structure database by hidden Markov models. Both Pfam and Gene3D are non-redundant (i.e. no overlapping), but can show repetitive domains within a protein.

Practical extraction and report language (Perl)

Practical Extraction and Report Language (Perl) is the most widely used programming language in bioinformatics, with its highly developed capacity to detect patterns, access and manipulate sequence and annotation data (http://www.perl.org). The majority of the automated work in this thesis was done by the scripts written in Perl, a high-level, general-purpose, interpreted, dynamic programming language. Some of the Perl scripts in this research are written in objected-oriented style. A huge collection of Perl modules, a discrete component of source code to be in a package, are open to public to save programmers' time and work. One of the most useful Perl toolkits is the BioPerl, a set of Perl modules built in an object-oriented manner for manipulating genomic and other biological data (http://www.bioperl.org) [8]. Bioperl provides an easy-to-use, stable, and consistent programming interface for biologist working on programming. To take a simple example, Swisskinfe, an object-oriented Perl library to handle Swiss-Prot entries, is very useful in parsing UniProtKB databases. Perl and BioPerl, with their open-source natures, will continue to provide the community automated analysis tools to deal with various biological issues.

Biology WorkBench and several bioinformatics tools

The Biology WorkBench is a web-based application integrating many bioinformatics tools (http://workbench.sdsc.edu/) [9]. This web server is designed to provide a comprehensive bioinformatics analysis environment, which could facilitate accessing and analysing the information. The interoperability between databases and programs could help researchers connect to multiple databases and analyze the data retrieved from different sources. Several bioinformatics tools often used in this research are listed as following:

- BLASTP: It compares a protein sequence to a protein database. Basic Local Alignment Search Tool (Blast), the most common tool in bioinformatics, utilizes the similarity among two sequences to suggest the closeness of functionality and homology.
- PSIBLAST: It is Position Specific Iterative version of Blast. It works in an iterative way, in which a scoring matrix is changed by related protein in each round, for the purpose of being more sensitive to find distant relatives of a protein query.

- CLUSTALW: It aligns several sequences at a time to compare multiple sequences. The result is derived by three steps: doing a pair-wise alignment, creating a phylogenetic tree, and achieving the multiple sequence alignment guided by the tree.
- TMHMM: It predicts the transmembrane helices in proteins with hidden Markov model. This prediction is especially important for membrane proteins in determining the topology of its structure.
- DRAWTREE: It draws an unrooted tree from a multiple sequence alignment, showing the inferred evolutionary relationship among several taxonomical units. The reconstruction of this evolutionary relationship provides a graphical understanding of the branching events along the evolution.
- DRAWGRAM: It is similar to DRAWTREE, but draws a rooted tree from a multiple sequence alignment. The major difference between them is the existence of most recent common ancestor of all taxonomical units in rooted tree.

Cytoscape

Cytoscape is an open-source software implemented for visualizing, analyzing and modelling of networks [10, 11]. In systems biology research, data is often integrated into networks, which shows the interactions among its components. In addition to its core functionality, Cytoscape is extensible through a plug-in manner, allowing versatile developments of additional computational analyses and features. It is of great power in the “omics” research fields when applied in conjunction with large databases, such as protein-protein, protein-DNA, and genetic interactions that are increasingly available. These are rich sources of information, describing the context and features of each interaction. As long as the data can be represented as nodes and edges, Cytoscape can display any kind of data as a network. This could help researchers exploring their studies in different aspects and facilitating the process of drawing knowledge from data. As the biological data grows rapidly and exponentially, the tools of analyzing networks become essential to many fields of biological researches.

References

1. Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res* 2011, 39:D214-219.

2. Magrane M, Consortium U: UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* 2011, 2011:bar009.

3. Mulder N, Apweiler R: InterPro and InterProScan: tools for protein sequence classification and comparison. *Methods Mol Biol* 2007, 396:59-70.

4. McDowall J, Hunter S: InterPro protein classification. *Methods Mol Biol* 2011, 694:37-47.

5. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al: InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009, 37:D211-215.

6. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, et al: The Pfam protein families database. *Nucleic Acids Res* 2010, 38:D211-222.

7. Lees J, Yeats C, Redfern O, Clegg A, Orengo C: Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Res* 2010, 38:D296-300.

8. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, et al: The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* 2002, 12:1611-1618.

9. Subramaniam S: The Biology Workbench--a seamless database and analysis environment for the biologist. *Proteins* 1998, 32:1-2.

10. Kohl M, Wiese S, Warscheid B: Cytoscape: software for visualization and analysis of biological networks. *Methods Mol Biol* 2011, 696:291-303.

11. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD: Cytoscape Web: an interactive web-based network browser. *Bioinformatics* 2010, 26:2347-2348.

CHAPTER 2:

DOMAIN-BASED IDENTIFICATION AND ANALYSIS OF GLUTAMATE RECEPTOR ION CHANNELS AND THEIR RELATIVES IN PROKARYOTES

Abstract

Voltage-gated and ligand-gated ion channels are used in eukaryotic organisms for the purpose of electrochemical signaling. There are prokaryotic homologues to major eukaryotic channels of these sorts, including voltage-gated sodium, potassium, and calcium channels, Ach-receptor and glutamate-receptor channels. The prokaryotic homologues have been less well characterized functionally than their eukaryotic counterparts.

In this study we identify likely prokaryotic functional counterparts of eukaryotic glutamate receptor channels by comprehensive analysis of the prokaryotic sequences in the context of known functional domains present in the eukaryotic members of this family. In particular, we searched the nonredundant protein database for all proteins containing the following motif: the two sections of the extracellular glutamate binding domain flanking two transmembrane helices. We discovered 100 prokaryotic sequences containing this motif, with a wide variety of functional annotations. Two groups within this family have the same topology as eukaryotic glutamate receptor channels. Group 1 has a potassium-like selectivity filter. Group 2 is most closely related to eukaryotic glutamate receptor channels. We present analysis of the functional domain architecture for the group of 100, a putative phylogenetic tree, comparison of the protein phylogeny

with the corresponding species phylogeny, consideration of the distribution of these proteins among classes of prokaryotes, and orthologous relationships between prokaryotic and human glutamate receptor channels. We introduce a construct called the Evolutionary Domain Network, which represents a putative pathway of domain rearrangements underlying the domain composition of present channels.

We believe that scientists interested in ion channels in general, and ligand-gated ion channels in particular, will be interested in this work. The work should also be of interest to bioinformatics researchers who are interested in the use of functional domain-based analysis in evolutionary and functional discovery.

Introduction

It is estimated that 20% - 40% of genes code for integral membrane proteins in archaea, bacteria, and eukaryote [1]. Because of the enormous energy barrier associated with moving ions across lipid bilayers [2] (Figure 2-1), proteins are essential for the transmembrane movement of polar and charged substances. Specific transmembrane proteins, like ion channels, transporters and pumps, appear to have arisen in very early forms of cellular life [3].

Ion channels are specialized transmembrane proteins through which cations or anions move passively down the electrochemical gradients that are created by ion pumps. Ion channels differ greatly in their structural and functional properties and are classified by their selectivity (Na^+, K^+, Ca^{2+} and Cl^-) and activation

mechanism (voltage-gated or ligand-gated). The largest subfamily of ion channels is comprised of the pore-loop channels, all of which carry a basic structural unit – a re-entrant pore-loop flanked by two transmembrane helices (TM's). (Figure 2-2) The ion selectivity is conferred by the pore-loop [4]. This common topology can be interpreted to suggest that the pore-loop channels have a common ancestor. This suggestion was born out by the discovery of a prokaryotic channel that contained the ligand-binding extracellular domain characteristic of glutamate receptor channels but a pore-loop characteristic of a potassium channel [5].

Glutamate, a major excitatory neurotransmitter, activates two receptor families: metabotropic glutamate receptor proteins (mGluR), which activate biochemical cascades, and ionotropic glutamate receptors, which form cation selective ion channels (iGluR) and are members of the pore-loop subfamily. Compared to the voltage-gated members of the pore-loop subfamily, iGluR's have opposite transmembrane orientation to the others (the pore-loop re-enters from the intracellular side). There are three major eukaryotic iGluR's subtypes, the AMPA, kainite and NMDA receptors, which form cation channels permeable to Na^+, K^+ and Ca^{2+}. Because of the difficulty of purification and crystallization of integral membrane proteins, we only have the high resolution structure for the extracellular ligand-binding domain of iGluR [6]. Some critical amino acids are identified in ligand-binding sequence.

In addition to the above-mentioned glutamate-receptor channel homologue, many other homologues to mammalian ion channels have been found in sequenced

prokaryotic genomes, such as K^+ channels, Na^+ channels, and Cl^- channels [7]. In addition Kuner, et al [8] noted the existence of other prokaryotic sequences bearing a resemblance to eukaryotic glutamate receptor channels.

The relative simplicity of prokaryotic ion channels makes them excellent objects for biophysical research [9]. A particularly notable example is the use of a prokaryotic potassium channel to make the first high resolution structure determination of voltage gated channels [10]. In many ways studying prokaryotic homologues can shed significant light on eukaryotic channels, as well the prokaryotic channels being of interest in their own right. For these reasons, a few years ago our laboratory (in collaboration with the laboratory of I. Aravind at NIH) set out to find prokaryotic homologues to the Ach receptor channel family. A straightforward BLAST [11] search yielded no results. We therefore undertook a search based on finding sequences with conserved domains characteristic of Ach receptor channel proteins and with the appropriate topology. That approach yielded a number of predicted prokaryotic members of this channel family [12]. One of our predicted channes was cloned, expressed, and functionally characterized as a channel [13] and high resolution structures were determined [14]. We anticipate that comprehensive identification of members of this group will lead to further functional and structural characterization of this family of channels, as well as insights into evolutionary and comparative aspects of channel biology. In the present study we extend this approach to a systematic domain-based search to identify and characterize in the nonredundant protein database all the prokaryotic homologues of the glutamate receptor channel family; i.e., prokaryotic iGluR's.

Materials and methods

Searching for prokaryotic iGluR's

The overall strategy for discovery of the prokaryotic iGluR's is provided in the flow chart of the five stage screening process, plus a validation stage using the InterPro database, in Figure 2-3(a).

We begin the search with the sequence iGluR0 from Synechocystic PCC6803 [15] which has been well characterized both functionally [5] and structurally [16]. At stage 1 in Figure 2-3(a), we used PSI-BLAST [11] to search the SDSC nonredundant protein database for the S1 binding region (NSEYVRQNSISAGITAVAEGELDILIGPISVTPERAAIEGITFTQPYFSSGIGLLIP, 57 aa long). This returned 2314 sequences with an E-value below 10. We applied the same method separately with the S2 segment of the binding region (EAVMFDRPALIYYTRQNPNLNLEVTEIRVSLEPYGFVLKENSPLQKTINVEMLNL LYSRVIAEFTERWL, 69 aa long) and returned 2344 sequences. At stage 2 in Figure 2-3(a), we invoked TransMembrane Hidden Markov Model [TMHMM] [17] to predict the number of transmembrane (TM) helices in each sequence. We eliminated all sequences with fewer than 2 TM's, which is the minimal number for the iGluR structure. This left us with 758 sequences with S1 and at least 2 TM's and with 731 sequences with S2 and at least 2 TM's. At stage 3, we separated the prokaryotic sequences from the eukaryotes. We found 135 sequences with S1 and 2 TM's and 132 sequences with S2 and 2 TM's. At stage 4, out of the

135 and the 132 we keep only the sequences that have both S1 and S2, which total 100. The annotations of the 100 sequences, clearly related to each other, are varied. In the definition line of the SDSC nonredundant protein database, 51 of them are annotated as ABC-type amino acid transporter or something similar, 13 of them are annotated as binding proteins, 14 of them are annotated as hypothetical proteins, 2 of them are annotated as K channels, plus some other scattered annotations (Table 2-1).

To explore the relationships among the 100 sequences, we aligned the sequences with ClustalW [18] and built a phylogenetic tree for them by DRAWGRAM [19]. The result is shown in Figure 2-4.

A notable feature of Figure 2-4 is that in many cases there is a disconnect between how close the sequences are on the tree and the similarity of the annotations. In some cases proteins that are quite similar are annotated differently, while sequences that seem quite far apart have the same annotation. A BLAST [11] of each of the 100 was done against the nonredundant database (data not shown) and confirmed that the sequence that gave the best hit was usually the one that was closest on the tree, and that the closest one on the tree was always one of the top few.

We then performed a topology analysis (stage 5 in Figure 2-3(a)) for the 100 sequences. The transmembrane regions are determined by TMHMM [17] and the glutamate binding regions are determined by sequence alignment. Through the visualization tool SeqVISTA [20], we can see the relative positions and lengths for TM's and glutamate binding regions in each protein. 22 of the 100 can be

identified as having the characteristic topology of glutamate receptor channels; i.e., the S1 and S2 glutamate binding domains flanking two TM helices (M1 and M2 region), in turn flanking a pore-loop (a domain that looks like a partial TM helix, P region). (One of the 22 sequences is the authoritative sequence that we used as our initial probe [5].) Figure 2-3(b) shows the e-values and TM probability scores for the S1/S2 and TM regions of the 22 sequences. It is seen that the statistical evidence for the identification and the topology are very strong. Figure 2-5 shows the SeqVISTA pattern characteristic of these 22 sequences and, for comparison, the SeqVISTA pattern for the human glutamate receptor channel orthologous (by the standard of reciprocal best hits) to the particular prokaryotic sequence shown. There are some differences. The human proteins are much larger, having an extra TM near the C-terminus. But there is a major similarity, i.e., the glutamate binding domains flanking two TM domains and a pore-loop. The supplementary material (Data not shown) includes the SeqVISTA diagrams for all 100 prokaryotic sequences in our search. Besides the 22 sequences, the other 78 prokaryotic sequences that have the glutamate binding domain and two or more TM helices have somewhat different topologies.

Results

Features and evolution of the prokaryotic glutamate receptor channels

Of the 22 putative channels, 12 of them have a distinctive potassium channel selectivity filter. We designate these as our Group 1. The other 10 have P

regions we do not recognize as distinctively similar to any channel with a known particular selectivity. Their annotations in the SDSC nonredundant protein database are shown in Table 2-2. Based on our analysis we would suggest that Group 1 be annotated as “putative glutamate-sensitive potassium channel” (except for #56, for which the word “putative” should be left off, since it has been functionally characterized as a glutamate-sensitive potassium channel [5].) We would suggest that Group 2 be annotated as “putative glutamate-sensitive ion channel”. Besides TM, we also used signalP [21] to test the existence of signal peptide. We found that two members of Group1 and two members of Group 2 lack the signal peptides which help the orientation of ion channel. The reasons for this may be the following: 1) They are pseudogenes; 2) they may have a different mechanism of inserting into membranes, or 3) they are oppositely oriented in the membrane than the other Group 1 and Group 2 channels. Motif searching has important significance in predicting the structures and functions of proteins. Therefore, we analyze the protein sequences by InterProScan [22] which is a web-based motif searching tool (http://www.ebi.ac.uk/interpro/) and federates 13 InterPro member databases into one resource. By searching the different protein signature databases, we can get a more comprehensive understanding of our target proteins. In order to efficiently utilize InterProScan, we developed a high throughput workflow around the InterProScan core program, that we call MotifNetwork [23].

Through MotifNetwork, we found that all 100 sequences have a glutamate binding motif, which was expected because we took glutamate binding region as our PSI-BLAST probe. We also found that none of the Group 1 or Group 2 members

had a domain characteristic of ABC transporters, reinforcing our view, stated above, that such annotation for those particular sequences is in error.

The results of the above are summarized in an Evolutionary Domain Network (EDN) (Figure2- 6). In the EDN representation, the proteins are grouped into domain sets according to the domain composition of each. (By “domain composistion” we mean the list of domains contained in the set.) The first row of the EDN contains all domain sets that consist of only a single domain. The second row contains those domain sets with two domains, the third row with three, etc. Tie lines are drawn between domain sets that can be derived from each other by the addition or subtraction of a single domain, representing roughly the evolutionary process of domain recombination. It should be noted that we have not screened out overlapping domains. Thus in some cases the same section of the protein sequence may be represented by two domain designations. We did attempt to screen overlaps, but any automated overlap screening resulted in loss of significant information, so we elected to report all MotifNetwork hits regardless of overlap.

By inspection of Figure 2-6, we see that all Group 1 sequences contain the IPR013099, whose short title is Ion transport 2. This domain represents a K^+ channel selectivity filter. As far as we have been able to determine so far, the combination of glutamate channel binding site and potassium channel selectivity filter represented by Group 1 is only in bacteria. No members of Group 1 can be found in archaea, neither can Group 2.

All Group 2 sequences have two domains in common: IPR001638 (Bacterial extracellular binding protein) and IPR015638 (glutamate receptor related). These are overlapping regions. The selectivity filter and permeation pathway have not apparently been defined as a distinctive InterPro domain.

Just one domain set appears disconnected from the others, and is placed on the right hand side of Figure 2-6. This contains domains IPR000515 and IPR013099. Only one protein (#94) is contained in this domain set. The existence of the potassium channel selectivity filter, plus the orientation of the glutamate binding domains to the transmembrane domains, defines this as a Group 1 channel. However the domain IPR000515, with this one exception, is only associated with the other sequences that do not have the structure of the glutamate binding domains flanking two TM domains and a pore-loop. It thus appears that sequence 94, despite its outlier status in Figure 2-6, may be a part of a linkage between the channel proteins and the non-channel proteins in this study. The intermediate domain sets have either vanished or have not yet been sequenced.

Inspection of Figure 2-6 shows that Human iGluR's can be connected to the prokaryotic scheme by intermediate steps equivalent to the net exchange of IPR001508 with IPR0016308 between NMDA receptor channels and Group 2 prokaryotic channels. This implies that Group 2 proteins might share a closer relationship to eukaryotic iGluR's than other prokaryotic glutamate-binding proteins and NMDA's are closer to prokaryotic iGluR's than are other eukaryotic iGluR's. Delta 1 protein reacquired IPR001638 (otherwises only found in prokaryotes among the group we are studying) in its motif composition, which may

result from a genetic recombination from outside (for example virus-mediated transfer from prokaryortes). It may be that some of the missing intermediates will appear in a more complete study of all the eukaryotic members of this family, which will be the subject of a future study.

Sequence analysis of group 1 and group 2 sequences

In order to identify the possible functions of Group 1 and Group 2 prokaryotic genes, we first made a multiple sequence alignment. In order to optimize the alignment, we align the domains separately and then join the alignments. We used the domain definitions of Mayer et al. [16] for the S1, S2, and channel regions (M1, P and M2). The conservation comparison is listed as Table 2-3. We can see that Group 2 is more conserved in glutamate binding region than Group 1 but less conserved in channel region.

In previous research about prokaryotic iGluR, scientists have identified some amino acids which are important in glutamate binding [5], specifically an Arg in S1 which interacts with α–carboxy group of L-glutamate and an Asp in S2 which interacts with α–amino group of L-glutamate. These are totally conserved in the Group 1 and Group 2 alignments. This conservation is shown in Figure 2-7.

Phylogenetic analysis of group 1 and group 2 sequences

We made phylogenetic trees for the different regions (S1, S2, and P region) in Group 1 and Group 2 sequences. It is seen that the trees have essentially the same structure. We can conclude that the glutamate binding region and channel region have remained together for a long time in evolutionary history.

We compared the phylogenetic tree of 16s rRNA genes with the phylogenetic tree of Group 1 and Group 2 genes in Figure 2-8. In this figure it is seen that in the tree of protein sequences (right hand tree) the Group 1 sequences (red) are clearly clustered together and separate from the Group 2 sequences (green). However in the 16s RNA sequences, the organisms containing Group 1 and Group 2 do not separate into distinct clusters from each other, indicating horizontal gene transfer (HGT) between the ancestors of some proteobacteria and some cyanobacteria.

Comparison with eukaryotic glutamate receptor channels.

Although iGluR research started with higher eukaryotic genomes, we still want to know if we can find all eukaryotic iGluR's by Group 1 and Group 2 sequences. First, we build a human iGluR list as a comparison by keyword search (Table 2-4).

Then, we used each of the Group1 and Group 2 as probes to blast human genome (BLASTP) [11], and accepted all hit with an e-value lower than 10. From the result (Table 2-5), we found that we can retrieve more human iGluR's using Group 2 as a probe. This implies that Group 2 sequences are closer to eukaryotic homologues than Group 1 sequences.

We also tested the orthologous relationship between eukaryotic iGluR prokaryotic iGluR by the "reciprocal-best-hits" criterion (data not shown). Both Group1 and Group2 members are orthologous to eukaryotic iGluR. This suggests two possible hypotheses. The first one is that Group 2 is the descendant of Group1 and eukaryotic iGluR is descendant of Group 2, because Group 2 is closer to

eukaryotic iGluR in the phylogenetic map (data not shown). The other hypothesis is that eukaryotic iGluR is descendant of Group 2 and Group 1 is the combination of Group 2 and prokaryotic potassium channels.

Discussions

Our results have implications for gene annotation, microbial communication and the evolution of cellular communication, and the origin and evolution of circadian rhythms.

Gene Annotation

The gene products we identified as being homologous to ionotropic glutamate receptors are largely annotated otherwise. In this paper, we did individualized analysis to identify these gene products as likely ionotropic glutamate receptors. The key addition to the previous annotation comes from analysis by functional domains and by how those domains fit into the overall topology of the protein, especially where they are relative to the transmembrane helices. Our group has developed a high-throughput computational environment for such scanning (MotifNetwork) [23], based on the functional domain definitions in the InterPro database. MotifNetwork is being enhanced to consider topology as well, so we anticipate that the procedures described in this paper will ultimately be completely automated.

Microbial Communication and the Evolution of Cellular Communication

In previous work our group used domain analysis to discover previously unknown prokaryotic members of the Ach Receptor Ion Channel family [12], a discovery which was later experimentally confirmed [13]. In this paper we extend the work to another major group of ligand-gated channels, the glutamate receptor channel family. These two discoveries together contribute to larger questions. What is the evolutionary origin of the electrochemical signaling mechanisms utilized in neuronal, neuromuscular, and neuroendocrine systems? To what extent do contemporary prokaryotes use these mechanisms to communicate? It should be noted that the patterns of occurrence of the two families of ligand-gated channels are very different. The prokaryotic Ach receptor channels are distributed across widely varying types of prokaryotes, both bacteria and archaea. By contrast, we found glutamate receptor channels only in bacteria, and clustered in particular bacterial subgroups. Because the sequence coverage of microbial genomes is still so sparse relative to the full range of microbial diversity, it is not possible to assess the full significance of this contrast. Based on our analysis of the existing data, it appears that horizontal transfer was the major mechanism for disseminating the prokaryotic members of the Ach receptor channel family. The members of the glutamate receptor channel family show evidence of at least two incidents of horizontal transfer (see Figure 2-8) but otherwise disseminate and variegate by descent. Based on the evolutionary domain network of the prokaryotic channels, it appears that domain reorganization was a significant factor in their evolution.

Origin and Evolution of Circadian Rhythms

We note three facts:

1) Among all prokaryotes, cyanobacteria have been shown to exhibit circadian rhythms [24].

2) In this paper, we find that among prokaryotes, ionotropic glutamate receptor channels are disproportionately present in cyanobacteria.

3) In animal brain slice preparations, glutamate resets circadian rhythms in a manner similar to light [25].

From this combination of facts, we are moved to suggest that glutamate signaling may provide a link connecting the circadian regulation of animals and cyanobacteria. This suggestion needs to be tested by further work.

References

1. Stevens TJ, Arkin IT (2000) Do more complex organisms have a greater proportion of membrane proteins in their genomes? Proteins 39: 417-420.
2. Gouaux E, Mackinnon R (2005) Principles of selective ion transport in channels and pumps. Science 310: 1461-1465.
3. Pohorille A, Schweighofer K, Wilson MA (2005) The origin and early evolution of membrane channels. Astrobiology 5: 1-17.
4. Ashcroft FM (2006) From molecule to malady. Nature 440: 440-447.

5. Chen GQ, Cui C, Mayer ML, Gouaux E (1999) Functional characterization of a potassium-selective prokaryotic glutamate receptor. Nature 402: 817-821.

6. Mayer ML (2006) Glutamate receptors at atomic resolution. Nature 440: 456-462.

7. Booth IR, Edwards MD, Miller S (2003) Bacterial ion channels. Biochemistry 42: 10045-10053.

8. Kuner T, Seeburg PH, Guy HR (2003) A common architecture for K+ channels and ionotropic glutamate receptors? Trends Neurosci 26: 27-32.

9. Kung C, Blount P (2004) Channels in microbes: so many holes to fill. Mol Microbiol 53: 373-380.

10. Doyle DA, Morais Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, et al. (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. Science 280: 69-77.

11. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

12. Tasneem A, Iyer LM, Jakobsson E, Aravind L (2005) Identification of the prokaryotic ligand-gated ion channels and their implications for the mechanisms and origins of animal Cys-loop ion channels. Genome Biol 6: R4.

13. Bocquet N, Prado de Carvalho L, Cartaud J, Neyton J, Le Poupon C, et al. (2007) A prokaryotic proton-gated ion channel from the nicotinic acetylcholine receptor family. Nature 445: 116-119.

14. Hilf RJ, Dutzler R (2008) X-ray structure of a prokaryotic pentameric ligand-gated ion channel. Nature 452: 375-379.

15. Saier MH, Jr. (2000) A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 64: 354-411.
16. Mayer ML, Olson R, Gouaux E (2001) Mechanisms for ligand binding to GluR0 ion channels: crystal structures of the glutamate and serine complexes and a closed apo state. J Mol Biol 311: 815-836.
17. Moller S, Croning MD, Apweiler R (2001) Evaluation of methods for the prediction of membrane spanning regions. Bioinformatics 17: 646-653.
18. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.
19. Felsenstein J (1993) PHYLIP (Phylogeny Inference Package). 3.5c ed.
20. Hu Z, Frith M, Niu T, Weng Z (2003) SeqVISTA: a graphical tool for sequence feature visualization and comparison. BMC Bioinformatics 4: 1.
21. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783-795.
22. Zdobnov EM, Apweiler R (2001) InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847-848.
23. Tilson JL, Rendon G, Ger M-F, Jakobsson E (2007) MotifNetwork: A Grid-enabled Workflow for High-throughput Domain Analysis of Biological Sequences. Bioinformatics and Bioengineering, 2007 BIBE 2007 Proceedings of the 7th IEEE International Conference pp. 620-627.
24. Mackey SR, Golden SS (2007) Winding up the cyanobacterial circadian clock. Trends Microbiol 15: 381-388.

25. Kalsbeek A, Kreier F, Fliers E, Sauerwein HP, Romijn JA, et al. (2007) Minireview: Circadian control of metabolism by the suprachiasmatic nuclei. Endocrinology 148: 5635-5639.

Tables

Table 2-1. Annotation of 100 bacterial sequences found to contain glutamate binding domains and two transmembrane domains.

gene annotation	protein No.	quantity
ABC transport system glutamine-binding protein	1,5,7,15,18,31,36,39,53,58,59,61	12
ABC-type amino acid transport/signal	2,4,6,8,10,12,13,17,24,26,27,28, 29,35,38,45,47,48, 49,54,55,64,69,71,77, 78,81,82,83,85,88,90,91,95, 96,97,98, 99,52	39
transporter	19,21	2
binding protein	66,80	2
extracellular solute-binding protein	9,25,33,40,41,46,63,65,70,73,89	11
hypothetical protein	16,20,22,43,50,56,57,60,67,72,7 5,76, 79,100	14
iGluR	3,23,37,62	4
K channel	42,94	2
sensory transduction protein kinase	34	1
sensory box protein	87	1
IMP dehydrogenase/GMP reductase	84	1
Unknown function	30,32,44,51,68,74,86,92,93	9

Table 2-2. Gene list of Group 1 and Group 2.

	Group 1	Protein ID
23	Possible ligand gated channel (GIC family	NP_896860.1
25	extracellular solute-binding protein, family	ZP_00674117.1
33	extracellular solute-binding protein, family 3	YP_378562.1
37	Ionoropic glutamate receptor	YP_376778.1
40	extracellular solute-binding protein, family 3	ABB23418.1
41	extracellular solute-binding protein, family	ZP_00517290.1
43	conserved protein of unknown function_ putative	YP_339120.1
46	extracellular solute-binding protein, family 3	ZP_00660701.1
56	hypothetical protein	NP_441171.1
62	Possible ligand gated channel (GIC family)	NP_894348.1
65	extracellular solute-binding protein, family	ZP_00530895.1
94	K channel, pore region	ZP_00533070.1
	Group 2	
1	ABC transport system glutamine-binding protein	NP_486951.1
2	COG0834: ABC-type amino acid transport/signal	ZP_00157839.2
3	Q3MEH3) Ionotropic glutamate receptor precursor	ABA20613.1
4	COG0834: ABC-type amino acid transport/signal	ZP_00108493.1
5	glutamine ABC transporter, periplasmic	YP_168531.1
6	COG0834: ABC-type amino acid transport/signal	ZP_00053934.2
9	extracellular solute-binding protein, family 3	ZP_00622239.1
42	glutamate-gated potassium channel	YP_204476.1
50	hypothetical protein	YP_132561.1
63	extracellular solute-binding protein, family 3	ZP_00629025.1

Table 2-3. Conservation comparison of Group 1 and Group 2.

		S1	S2	channel
Group 1	identical	10/97	1/132	17/115
Group 1	Strongly conserved	10/97	15/132	25/115
Group 1	Weakly conserved	9/97	12/132	10/115
Group 2	Identical	12/93	10/129	2/120
Group 2	Strongly conserved	16/93	17/129	15/120
Group 2	Weakly conserved	6/93	11/129	11/120

Table 2-4. Human iGluR's.

AMPA	AMPA 1	NP_000818.1	906 aa
AMPA	AMPA 2 isoform 1	NP_000817.2	883 aa
AMPA	AMPA 2 isoform 2	NP_001077088.1	883 aa
AMPA	AMPA 2 isoform 3	NP_001077089.1	836 aa
AMPA	glutamate receptor 3 isoform flip	NP_015564.4	894 aa
AMPA	glutamate receptor 3 isoform flop	NP_000819.3	894 aa
AMPA	AMPA 4 isoform 1	NP_000820.3	902 aa
AMPA	AMPA 4 isoform 2	NP_001070711.2	884 aa
Kainate	kainate 1 isoform 1	NP_000821.1	918 aa
Kainate	kainate 1 isoform 2	NP_783300.1	905 aa
Kainate	kainate 2 isoform 1	NP_068775.1	908 aa
Kainate	kainate 2 isoform 2	NP_786944.1	869 aa
Kainate	kainite 3	NP_000822.2	919 aa
Kainate	glutamate receptor KA1	NP_055434.2	956 aa
Kainate	glutamate receptor KA2	NP_002079.3	980 aa
NMDA	NMDA receptor 1 isoform NR1-1	NP_000823.4	885 aa
NMDA	NMDA receptor 1 isoform NR1-2	NP_067544.1	901 aa
NMDA	NMDA receptor 1 isoform NR1-3	NP_015566.1	938 aa
NMDA	N-methyl-D-aspartate receptor subunit 2A	NP_000824.1	1464 aa
NMDA	N-methyl-D-aspartate receptor subunit 2D	NP_000825.2	1336 aa
NMDA	N-methyl-D-aspartate receptor subunit 2C	NP_000826.2	1233 aa
NMDA	N-methyl-D-aspartate receptor subunit 2B	NP_000827.2	1484 aa
NMDA	N-methyl-D-aspartate 3A	NP_597702.1	1115 aa
NMDA	N-methyl-D-aspartate 3B	NP_619635.1	1043 aa
Delta	delta 1	NP_060021.1	1009 aa
Delta	delta 2	NP_001501.2	1007 aa

Table 2-5. Reverse BLAST result against human genome using Group 1 and Group2 as a probe.

Group 1		Group 2	
protein No.	ratio	protein No.	ratio
23	13/26	1	26/26
25	26/26	2	26/26
33	26/26	3	26/26
37	13/26	4	26/26
40	26/26	5	26/26
41	24/26	6	25/26
43	19/26	9	26/26
46	26/26	42	25/26
56	16/26	50	20/26
62	14/26	63	18/26
65	26/26		
94	26/26		

Figures

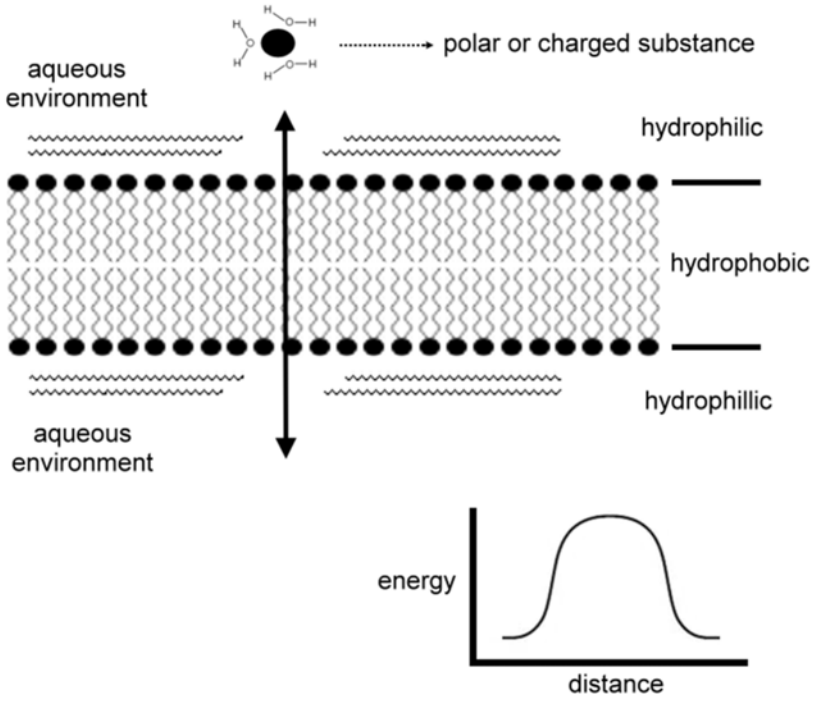


Figure 2-1. - Energy barrier for an ion to move across the membrane.

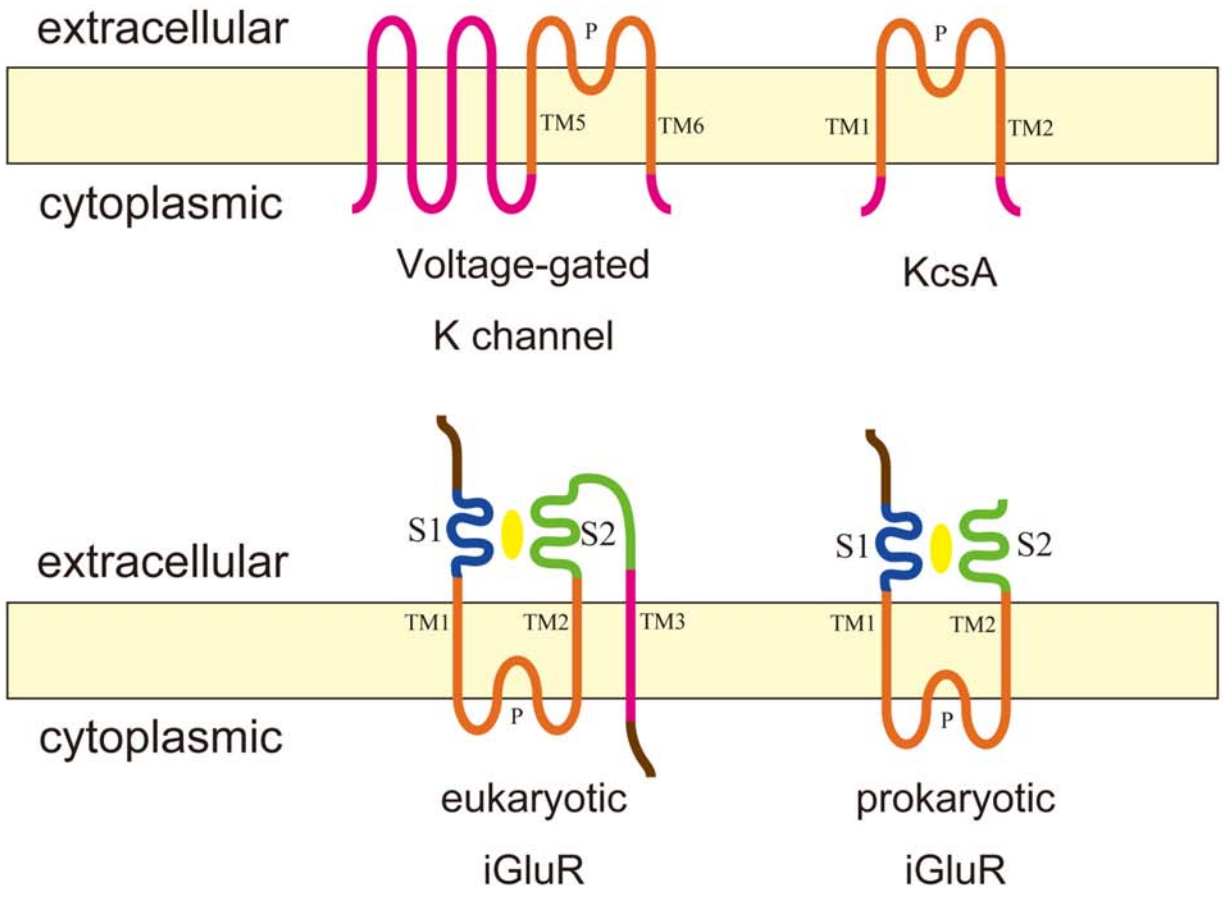
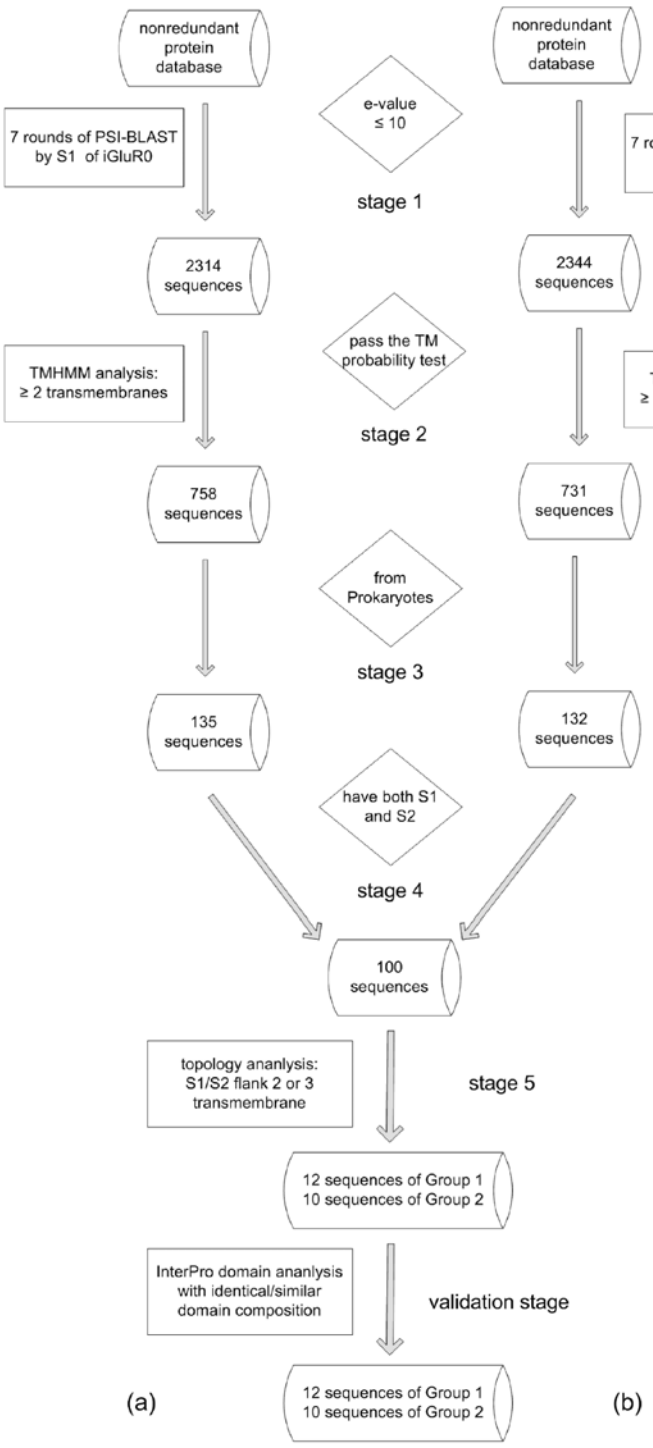


Figure 2-2. - Topology diagrams of pore-loop ion channels.



#	E-value for S1	E-value for S2	TM probability for M1	TM probability for P	TM probability for M2
Group 1					
23	3.00E-07	1.00E-09	0.99999	0.89394	0.99275
25	9.00E-08	2.00E-09	1	0.99337	0.99813
33	0.018	8.00E-09	0.99999	0.73089	0.99722
37	1.00E-08	3.00E-08	1	0.99092	0.99848
40	1.00E-07	4.00E-08	0.99996	0.96695	0.99982
41	6.00E-05	6.00E-08	0.99997	0.97498	0.99693
43	1.00E-06	7.00E-08	1	0.69016	0.99812
46	3.00E-07	1.00E-07	0.99975	0.70932	0.99858
56	6.00E-10	6.00E-07	0.99984	0.06564	0.99552
62	2.00E-10	1.00E-06	0.99993	0.88389	0.99434
65	6.00E-04	3.00E-06	1	0.77978	0.99838
94	2.00E-05	2.1	1	0.43197	0.99847
Group 2					
1	5.00E-07	6.00E-16	0.99999	0.45822	0.99995
2	2.00E-06	7.00E-16	0.99998	0.47037	0.99995
3	1.00E-06	9.00E-16	0.99999	0.47272	0.99995
4	1.00E-06	1.00E-15	0.99998	0.3648	0.99995
5	2.00E-08	1.00E-15	0.99998	0.31253	0.99836
6	2.00E-06	1.00E-13	0.99924	0.95602	0.99996
9	3.00E-06	9.00E-13	0.99999	0.19217	0.99835
42	1.00E-03	6.00E-08	0.99988	0.78611	0.99969
50	2.00E-04	2.00E-07	0.99999	0.95594	0.99994
63	4.00E-07	1.00E-06	0.9989	0.02518	0.99825

Figure 2-3. - The searching strategy for finding prokaryotic iGluR□s and the statistical proof.
The strategy in (a) includes 5 stages and an additional validation stage. At each stage, we select protein sequences which are qualified for the requirements. In (b), the statistical e-values for S1 and S2 identification and TM probability scores by PSI-BLAST and TMHMM, respectively. The TM probability scores which do not pass the TM probability test are shaded in (b) and not counted as TM□S.

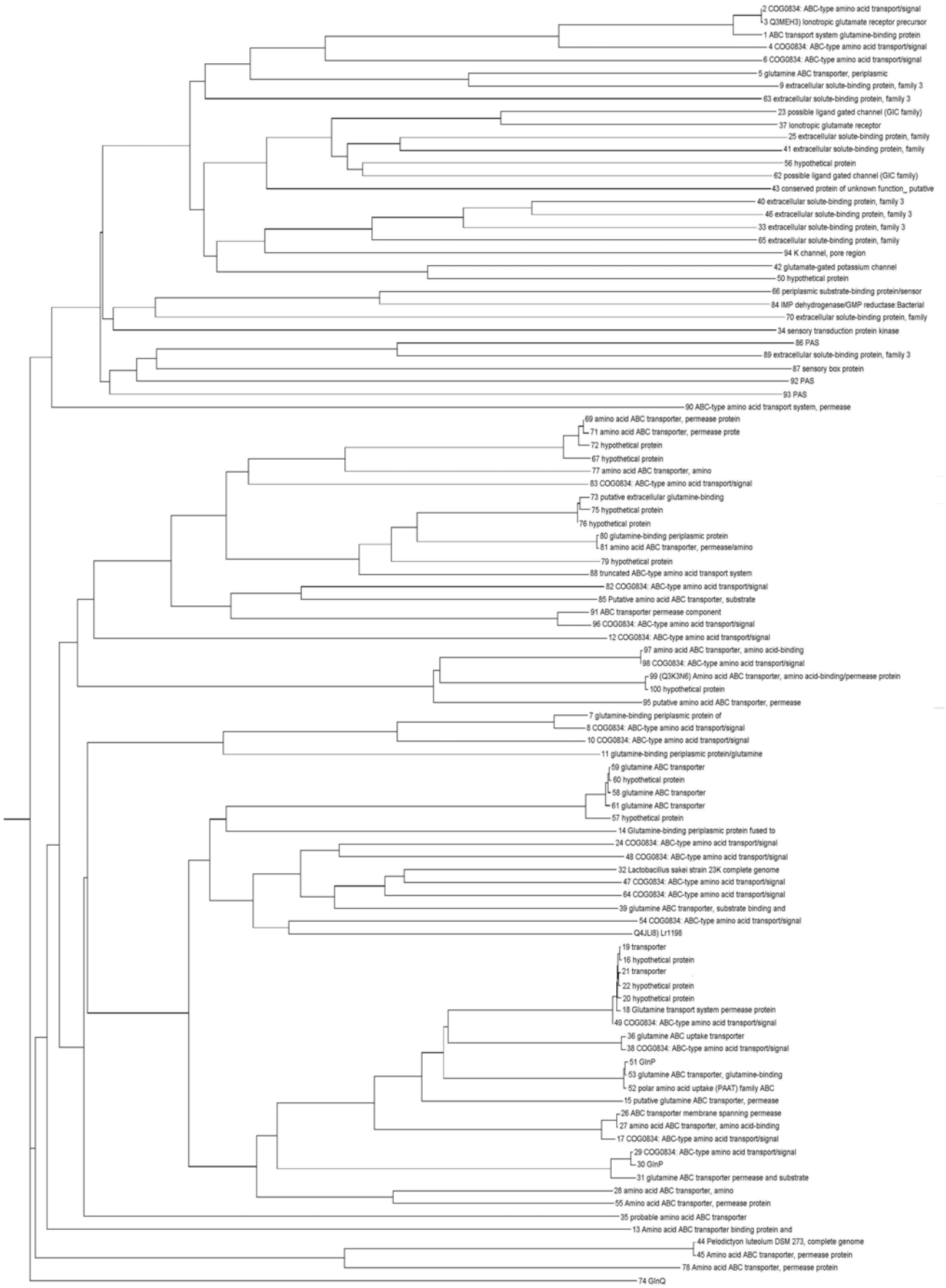


Figure 2-4. - Phylogenetic tree for 100 sequences.
Phylogenetic tree for 100 potential prokaryotic glutamate receptor channels as determined by presence of glutamate binding domain and transmembrane helices. The sequences are labeled with the definition line from the SDSC nonredundant protein database.

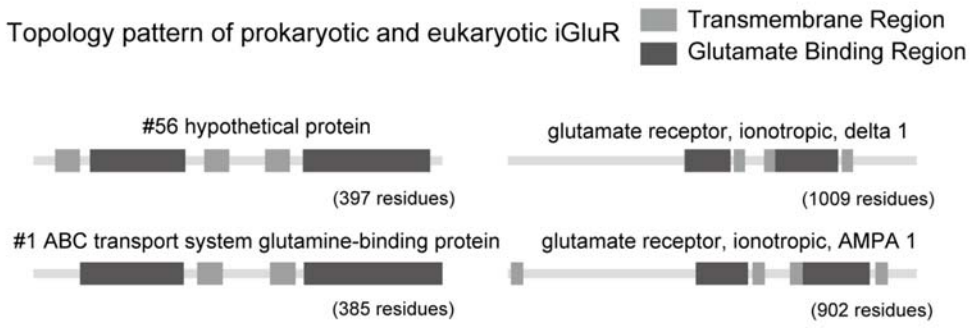


Figure 2-5. - Topology pattern for Group 1 and Group 2.
The eukaryotic counterpart of prokaryotic iGluR #56 is delta 1. The eukaryotic counterpart of prokaryotic iGluR #1 is AMPA 1. They are all with the structure of the S1 and S2 glutamate binding domains flanking two TM helices, in turn flanking a P-loop.

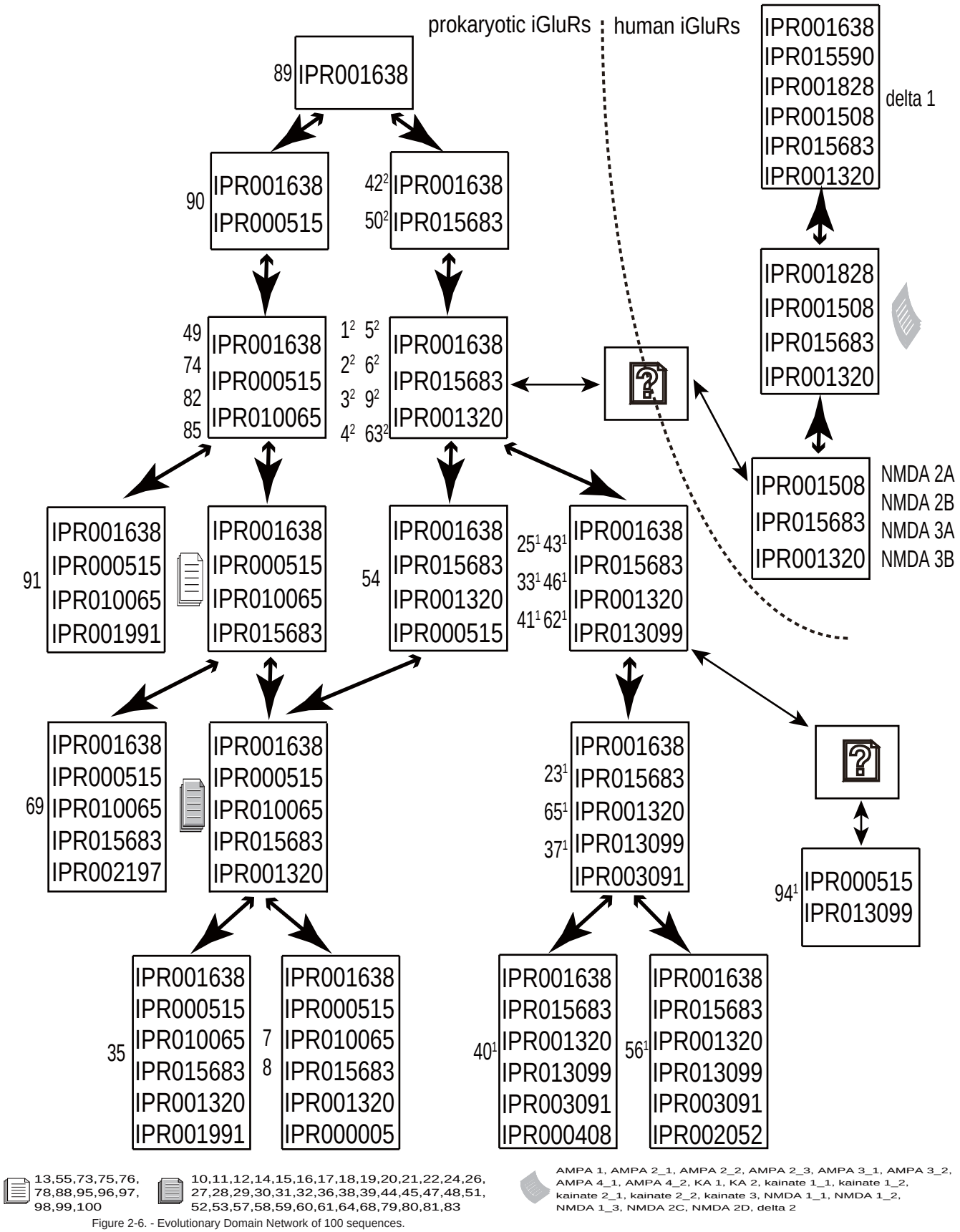


Figure 2-6. - Evolutionary Domain Network of 100 sequences.

IPR001638: Bacterial extracellular solute-binding protein, family 3. IPR015683: Glutamate receptor-related. IPR000515: Binding-protein-dependent transport systems inner membrane component. IPR010065: Amino acid ABC transporter, permease protein, 3-TM region, His/Glu/Gln/Arg/opine. IPR001320: Ionotropic glutamate receptor. IPR013099: Ion transport 2. IPR003091: Voltage-dependent potassium channel. IPR001991: Sodium:dicarboxylate symporter. IPR002197: Helix-turn-helix, Fis-type. IPR000005: Helix-turn-helix, AraC type. IPR000408: Regulator of chromosome condensation, RCC1. IPR002052: N-6 Adenine-specific DNA methylase. IPR001508: NMDA receptor. IPR001828: Extracellular ligand-binding receptor. IPR015590: Aldehyde dehydrogenase.

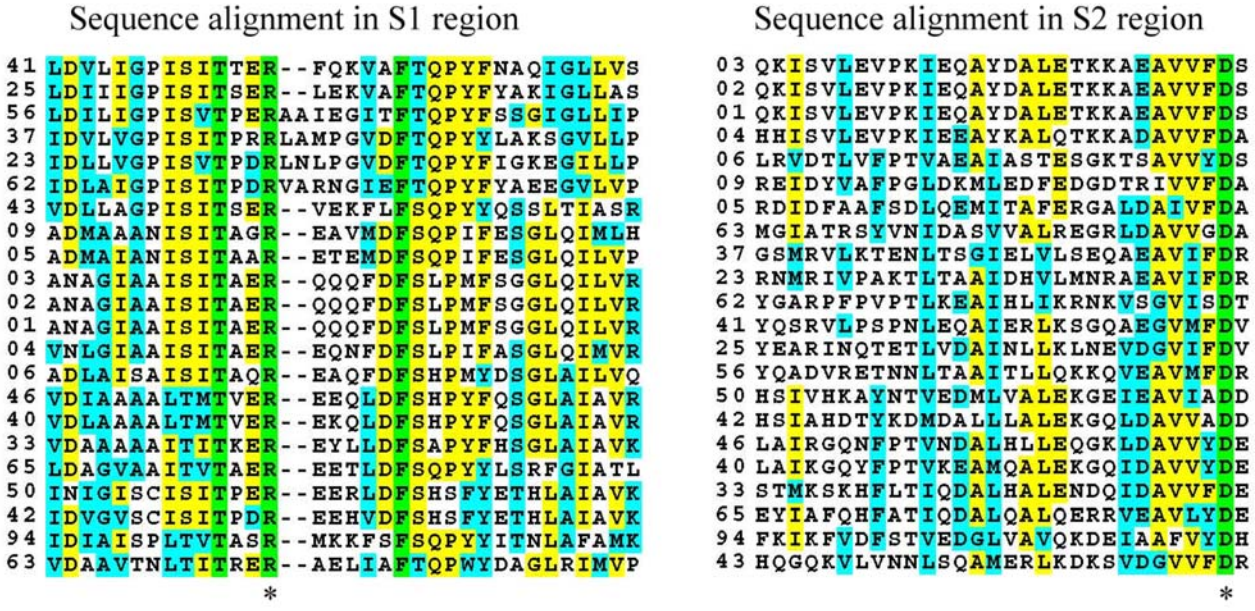


Figure 2-7. - The sequence alignment in S1 and S2 region of Group1 and Group 2 proteins.

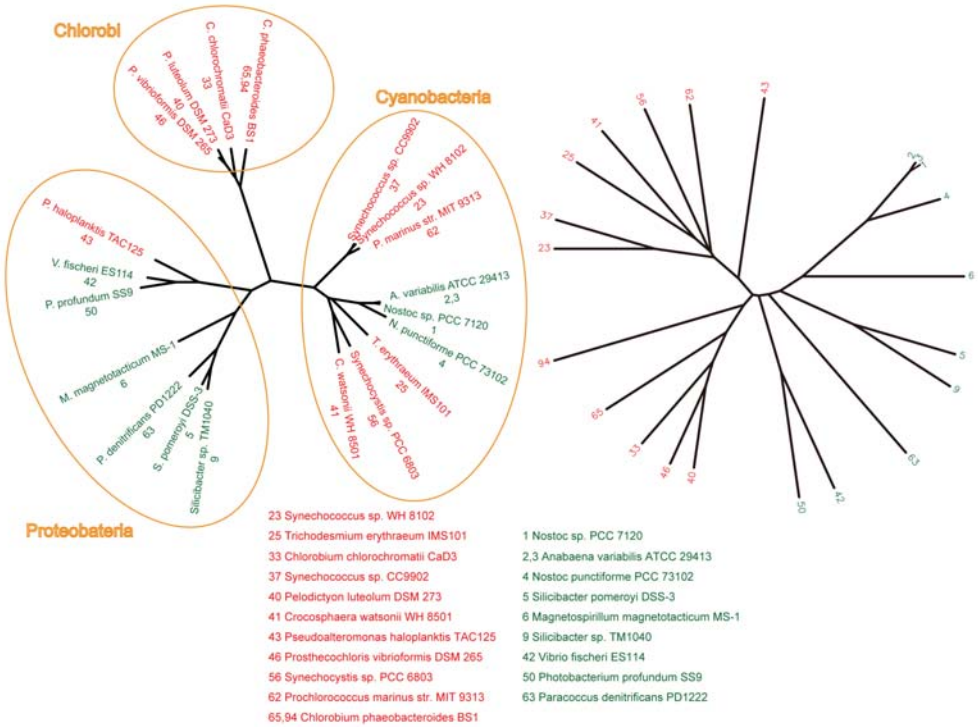


Figure 2-8. - Phylogenetic trees of 16s rRNA genes and Group 1/ Group 2 genes.
Left hand side is the 16sRNA tree for the species that contain Group 1 and Group 2 prokaryotic glutamate receptor channels. Right hand side is the tree for the Group 1 and Group 2 proteins. The fact that the clustering patterns are different for the two trees indicates horizontal gene transfer of glutamate receptor channels among the bacteria. In particular, it seems there must have been a minimum of two transfers, one from cyanobacteria to proteobacteria, and one from proteobacteria to cyanobacteria.

CHAPTER 3:

EVOLUTION OF DOMAIN COMPLEXITY IN PROTEINS

Abstract

In this paper domain databases are segmented into biological taxa to explore the evolution of domain complexity over evolutionary time. A new finding, in contrast to previous studies, is that by domain abundance, bacterial and archaeal proteins are as complex as eukaryotic proteins. Within all three superkingdoms there is a trend that the proteins that are unique to one superkingdom have more domains per protein than proteins that are shared between two or three superkingdoms. On the other hand, protein domains that are shared between superkingdoms are more “promiscuous”; i.e., they appear in combination with more other domains than domains that are unique to one superkingdom. By these measures of complexity, the issue of early emergence in evolutionary history (as measured by degree of common occurrence across the superkingdoms) is a more important determinant of complexity than the issue of which of the three superkingdoms a particular domain or domain combination appears.

Background

Among the imperative objectives in molecular biology is to resolve the functions and structures of proteins which are voluminously emerging through

sequencing projects. Homology identification, which can be determined by a method based on the entire protein sequence or a method based on the domain content of the protein, usually lies at the first step of achieving the above goal. The sequence-based approach generally gives good results for proteins from closely related organisms. For more distantly related sequences, the domain-based approach may be more effective, because it is able to account for the rearranging of domains within the protein that occurs over long evolutionary times[1].

Domains are not only the basic units of function and structure, but also the building blocks of proteins through evolution events. Although the word “domain” has been used extensively in molecular biology with slightly different definitions, it is commonly accepted that a domain is a compact, spatially distinct unit which usually folds independently of other domains and shares conserved sequence with homologous proteins. In the course of evolution, nature tends to reuse and recombine existing modules to expand the versatility of proteins, in addition to sometimes inventing new modules [2, 3]. Therefore, the arrangement of domains can be explored to understand the evolutionary process. Major contributing modes of domain rearrangement are duplication, divergence, recombination, fission, and fusion [4].

The methods of studying proteins at the domain level are different from the methods at the sequence level. Instead of viewing proteins as a sequence of amino acids, the domain perspective views proteins as domain compositions (a collection of domains) or the domain architectures (a sequential order of domains). Domain-based homology identification has been proved to be a

sensitive way to find common functionalities between distantly related proteins and proteins in distantly related organisms [5, 6]. By comparing domain composition or domain architecture, similarity measurement can be determined to evaluate the evolutionary distance between distantly related as well as closely related proteins [1, 7-9]. This information can be used to infer function and establish evolution history. Protein-protein evolutionary networks can be constructed by defining the nodes as proteins and edges as sharing common domain(s) between proteins (i.e. having a non-zero similarity score). This type of network provides a new way to investigate the proteome in a large scale [6].

The phenomenon of biological complexity is of fundamental interest to scientists. Complexity of regulatory networks and the functional versatility of proteins are both useful measures of complexity, whereas the total number of genes in an organism does not necessarily correlate well with functional complexity [10]. Eukaryotes have been reported to have more multi-domain proteins than prokaryotes [11, 12]. In this paper we will revisit the issue of the relative complexity of proteins in eukaryotes, bacteria, and archaea using comprehensive analysis of protein and domain databases.

The capability of a domain to form different domain compositions/domain architecture is termed promiscuity (or mobility) [8, 13, 14]. Among several methods to measure the promiscuity, co-occurrence of two domains is a simple but accurate way [15]. These promiscuous domains are involved in many processes of protein-protein interaction in eukaryotes, such as signal transduction [8]. A domain-domain co-occurrence can be constructed by

defining nodes as domains and edges as co-occurrence within at least one protein. By analyzing this type of network, an early study found that promiscuous domains form “hubs” with high degrees of connectivity and the network is approximately scale-free [15].

Up to date, there are a large variety of domain databases, each constructed in fundamentally different ways. The methods of domain definitions include sequence clustering (e.g. ProDom), regular expression (e.g. PROSITE), profiles (e.g. PROSITE, HAMAP, PRINTS), and Hidden Markov Models (e.g. Pfam, SMART, Gene3D) [16]. Gene3D domain definition is based on structural information and derived from CATH database, a hierarchical classification of protein domain structures [17]. By contrast, Pfam domain definition is based on sequence and derived from multiple sequence alignment [18]. Additionally, InterPro is an integrated domain information database of 11 protein signature databases which use different methods for defining signature, including sequence clustering, regular expression, profiles and hidden Markov models [16, 19]. InterPro database provides a more comprehensive (albeit sometimes redundant) coverage than any one of its constituent databases. It creates a unique InterPro Record (IPR) representing a specific domain signature, which in turn can be used to identify unknown sequences. InterPro database has been applied in the automatic annotation of the UniProtKB/TrEMBL [20].

In this research, we analyzed the domains and domain combinations from InterPro, Pfam, and Gene3D domain definitions. First, the domain connectivity was investigated at the superkingdom level, and then followed by

the numbers of domains per protein. These two factors represent the complexity of domain content and its implication in biological complexity was investigated as well. We also extended the similar analyses to proteomes of some eukaryotic species to understand if domain content has implication in organismic complexity.

Results

Domain content retrieval

We started by parsing the SWISS-Prot database and the TrEMBL database in the UniProtKB v.2011_01. The information of InterPro domain, Pfam domain, Gene3D domain, and taxonomical data, were retrieved from 13,592,921 proteins (524,420 from Swiss-Prot and13,069,501 from TrEMBL; 62.25% from Bacteria, 1.90% from Archaea, 27.56% from Eukaryota). As shown in Table 3-1, the coverage for InterPro domain definition is highest (77.50% of UniProt proteins contain at least one InterPro domain), compared to Pfam domain definition (73.4%) and Gene3D domain definition (32.96%). The number of the domains and the domain compositions in InterPro domain definition is also larger than in the Pfam domain definition and much larger than Gene3D domain definition (21,091, 11,464 and 1,147 domains respectively), suggesting a more comprehensive functional classification by InterPro domain definition. While Gene3D domain definition is derived from structural conservation alone and Pfam is derived from the sequence conservation, InterPro domain definition is integrated from 11 domain databases including Pfam and Gene3D,

so these results are as expected. For each of the domain definitions, the domains were incorporated into the domain-domain co-occurrence networks and the degree of connectivity (number of different domains in which it co-occurs in at least one protein) for each domain was counted (See Methods). The average degree of connectivity is higher in InterPro domain definition (23.10) than in Pfam domain definition (7.71) and Gene3D domain definition (9.92). One reason for the high connectivity in the InterPro domain definition is apparently that differently defined InterPro domains often overlap each other, so connections between domains are sometimes essentially connections of overlapping domains. Among the proteins carrying domains defined by each domain definition, the number of distinct domain compositions were counted. By InterPro domain definition, there are 153,165 different domain compositions. The average number of domains within a protein is 2.77 by InterPro domain definition (proteins with no identified InterPro domains are excluded). The corresponding number is 1.48 by Pfam domain definition and 1.47 by Gene3D domain definition (again excluding proteins without domain definition). In the Pfam and Gene3D domain definition, repeats can either be counted (repetitious domains within the same protein are regarded as different) or not. Therefore, in Pfam there are 50,369 domain compositions (not counting repeats) or 77,561 domain compositions (counting repeats). There are 6,869 Gene3D domain compositions (not counting repeats) or 14,451 Gene3D domain compositions (counting repeats).

Domain-domain network and degree of connectivity

Domain-domain networks based on InterPro, Pfam and Gene3D definition were also constructed. In this type of network, nodes are the domains and edges are the co-occurrences of two domains within a protein. Degree of connectivity on each node is the number of edges connected to this node and sometimes is defined as the promiscuity or mobility of a domain, which shows the capability to form different domain compositions (See an example in Methods). As mentioned above, domains with high degree of connectivity (i.e., domains appearing in many different domain compositions) often are involved in protein-protein interactions and signal transduction pathways [8]. We investigated the distribution of degrees of connectivity in three domain definitions (Figure 3-1A). They all follow a power-law, which demonstrates that the co-occurrence does not happen randomly. It has been suggested that the power-law distribution is characteristic of robust and error-tolerant networks [21]. From Figure 3-1B, in which the degrees of connectivity were calculated by the taxonomy and the connections in the respective taxonomy were counted, it is suggested that Eukayotes have a higher average of degree of connectivity than either Bacteria or Archaea. The average in Archaea is lowest perhaps due to the fact that the domain annotation is less comprehensive in Archaea. These trends are similar in all three domain definitions.

We divided the domains into seven categories by their existences in major lineages (Bacteria, Archaea, and Eukaryota, Figure 3-1C). The seven categories include: unique to one of the three lineages (3 categories), shared

by two of the three lineages (3 categories) and shared by all three lineages (1 category). In this computation, all connections were counted. Within each category, the characteristic scale-free structure of the networks also holds (shown in supplementary data) in all three domain definitions. In Figure 3-1C, the numbers of the domains are shown on a Venn diagram. Points to consider when examining Figure 3-1C are

- Only about 1/5 of the InterPro and Pfam domains are common to all three superkingdoms, while over 40% of the Gene3D domains are common to all three. This discrepancy may be the result of experimental selection; i.e., a very favorable target for structure determination is a prokaryotic protein that has a functionally important eukaryotic counterpart.
- The domains common to all three superkingdoms presumably emerged earlier in evolutionary history than domains common to two superkingdoms or unique to one, and are engaged in the cellular activities universal to all living organisms. The domains unique to one superkingdom presumably appeared after the division of three superkingdoms and contribute the unique characteristics to each superkingdom.
- We calculated the average degrees of connectivity per domain and the standard deviations in each category. The fact that the standard deviations are much greater than the mean value indicates the existence of outliers; i.e., some domains that are connected to a very large number of other domains.
- The division into seven categories does reveal trends in domain

connectivity. Within each of the three kingdoms the connectivity is lowest for the domains unique to that kingdom, somewhat higher for domains shared with one of the other kingdoms, and highest for domains common to all three kingdoms.

- While the overall connectivity of eukaryotic domains is not significantly greater than that for archea and bacteria, the connectivity of the domains unique to eukaryotes is significantly greater than the connectivity of domains unique to archaea and bacteria. Perhaps this is correlated to what makes eukaryotes uniquely different from prokaryotes.

The analysis of domain compositions

Next, we investigated the distributions of the number of domains within a protein for each domain definition. In Figure 3-2A, we plotted the number of proteins displaying a given number of different domains for each of the domain definitions. In this figure, no matter how many times a domain appears in a protein, we count it as one domain. We see that the envelope of the histogram is an exponentially decreasing function. In Figure 3-2B, we count each the repeat as another domain. The distribution for InterPro domains continues to follow a simple exponential decay. However, the graphs for Pfam and Gene3D definitions show a distinctive departure from the simple exponential. These results are consistent with two distinct mechanisms for adding new domains on the one hand, or repeating a domain on the other hand. The former can be modeled by a simple first order process where individual domains have a fixed probability in evolutionary time for either being

split off, or being added to, a given domain composition, giving rise to the exponential relationship. Domain repeats, on the other hand, seem to follow a different dynamic, with repetitive domains being produced by internal duplication, while new domains are acquired from other proteins. Therefore, we redid the graph by counting repetitive domains as one in Figure 3-2B. The results of Pfam and Gene3D definition show the same pattern of distributions with InterPro definition, i.e. the exponential distribution. Therefore, it can be suggested that the dynamics of adding a new domain to a protein is different from repeating a existing domain in a protein. Note that our exponential distribution for number of domains vs. number of proteins is qualitatively different from the power-law distribution suggested in [22]. We infer that the difference lies in the more complete data available today than when the previous analysis was published.

We further categorized the domain compositions by lineages (Bacteria, Archaea, and Eukaryota) (Figure 3-3). Specifically, we categorized the proteins by the 7-way division (see Methods), according to which of the three superkingdoms the domain compositions appeared in. Then we calculated the average numbers of domains for the proteins in each category and in each superkingdom. The average numbers of domains per protein in InterPro domain database are similar in Bacteria, Archaea and Eukaryota (B: 2.81; A: 2.64, E: 2.77), and also in Pfam (B: 1.44; A: 1.37, E: 1.58) and Gene3D domain databases (B: 1.44; A: 1.36, E: 1.52). Note similarity of the trends for each of the databases, even though the domain definitions are different. This suggests that Eukaryota has no significantly higher domain complexity than Prokaryota, as measured by number of domains per protein using any of the

domain definitions—InterPro, Pfam, or Gene3D. The average number of domains by seven categories gives another perspective into complexity. The average number of domains in a domain set is lowest in among those domain sets found in all of the superkingdoms, somewhat higher in those domain sets shared between two superkingdoms, and higher yet for the domain sets unique to one superkingdom. It appears therefore that the pathway of evolution in each of the three superkingdoms has been in the direction of greater domain complexity. Domain complexity is more a function of being of recent origin than of being in any one of the superkingdoms.

The analysis of domain content in 24 eukaryotic proteomes

The similar analyses of the complexity in terms of domain were applied to some eukaryotic organisms. In Figure 3-4, 24 eukaryotes, which are completely sequenced and cover the major eukaryotic branches, are illustrated in their taxonomical hierarchy. All these 24 species can be grouped into animals, fungi, plants and protists. From Table 3-2, considering the average number of InterPro domains, protists has the slightly lower numbers and animals has the slightly higher numbers. Moreover, we plotted the distribution of domain numbers within each species (Figure 3-5). The curve of number of proteins with a given number of domains, vs. the number of domains, obeys an exponential relationship within each species. (Values listed in Table 3-3). The less negative the parameter is, the stronger the tendency of more domains in a protein is. Animals and fungi have slightly higher tendency than plants and protists. From averages and distributions of the numbers of domains in a protein, although the difference is observed, it may

not be sufficiently different to be a contributing factor to organismic complexity.

Secondly, as for the average number of degree of domain connectivity, there is also a slightly difference. In Table 3-2, animals still have the higher connectivity. However, that plants rank the second suggests that multi-cell organisms need to have higher connectivity in domain to fulfill the communication of proteins. The power-law distributions of domain connectivity within each species also hold. However, it is again not significant enough to confirm that this factor is one of the major contributing factors to organismic complexity, although degrees of connectivity is significant divergent at the superkingdom level. Domain abundance and connectivity may play another role in this, but other factors should be included.

Implications of combination of domains

As far as the hierarchy is considered, amino acids are the basal level of proteins and domains are at the second level, which can be used as the building units to construct domain sets, which can be regarded as the third level. Domain combination is clearly a major force in developing versatility of function in proteins. A potassium channel selectivity filter, for example, assumes importance in many different biological functions because it has been combined with regulatory domains which open and close the channel in response to membrane potential, calcium, g-proteins, pH, redox potential, etc. We do not yet have a mathematical model describing domain combination during evolution, which would include the processes of duplication, fusion of domains, fission of domains, extinction of domains and domain sets, etc. Such a model would need to explain, for example, the exponential relationship

between numbers of domains in domain sets vs. the size of the domain sets that is seen in each of the kingdoms in Figure 3-2. Even without such a model, the results in Figure 3-3 suggest to us that the long-term direction of protein evolution in all the kingdoms is in the direction of greater complexity. Because domain compositions represent the functionality of proteins, the number of domain compositions is an indication of the versatility of functionality. We calculated the capabilities of forming InterPro domain combinations. The result is in Figure 3-7 (See Methods for calculations). Archaea have produced a larger fraction of the domain combinations available to them based on the number of their domains than either Bacteria or Eukaryota. The domains common to all three superkingdoms have explored a larger fraction of the combinations available to them than those that are shared by any two kingdoms or unique to any single superkingdom (Figure 3-7A). The same analysis was done wihin 24 Eukaryotes (Figure 3-7B). Protists have relatively higher capabilities to create domain combinations from the available domains and animals have relatively lower capabilities. We infer that species with lower complexity develop a way to increase the adaption to environment by higher ability to create domain combinations, although the mechanism is not clear yet. The results of different database coverage (See Methods) show consistent results with Figure3-7B (data not shown).

Discussions

Domain abundance

Although the previous research showed that Eukaryota have higher percentage of multi-domain proteins [12, 22], our result show the opposite phenomenon that Eukaryota has similar domain abundance with Archaea and Bacteria. The difference reflects different methodologies. In our work we used all available domains as defined in the respective databases and only those domains. In [22] the authors counted regions between identifiable domains as domains themselves. We suggest this overstates the complexity of the protein, because the regions between identifiable domains are highly variable and therefore are of low complexity. We choose therefore not to count them. In [22] the authors use only a portion of Pfam (Pfam A) which contains domains for which there is experimental evidence of function. Pfam B (which is omitted from the analysis of [22]) contains initial domains inferred from statistics of conservation, regardless of whether a function has yet been identified. We believe this should be included, on the grounds that conservation implies adaptive value and hence function, whether or not the function has yet been identified. In summary, we included in our analysis all high-complexity (i.e., conserved) regions and only those regions, while [12] included low-complexity regions as well and [22]omitted some high complexity regions.

Our results suggest that domain abundance is not a factor which contributes to the functional complexities at superkingdom level or at species level. While the protein lengths are longer in Eukaryota, abundance of conserved domains

does not increase with it. It may be that functional complexity is more related to the interactions a protein can provide than the domain complexity a protein could embody.

In the result of domain abundance 7-way division, it is suggested that the increase of domain abundance is a later event in the history of life. Proteins with the domain compositions shared by three superkingdoms have the relatively lower domain abundance than the protein with the domain compositions share by two superkingdoms. Proteins with the domain compositions unique to a single superkingdom have the highest domain abundance. Since the domain combinations unique to each of the superkingdoms represent those that have arisen most recently in evolutionary history, this supports the thesis that the direction of evolution is towards greater complexity [23]. The might imply that domain abundance is a way to enhance the adaption to environment, but not the domain complexity.

Similarly, in the result of 24 eukaryotes, the result of domain abundance shows no significant difference. This reinforces the above conclusion frown from the result of 3-way division.

Domain connectivity

We find that domains in eukaryotes have greater promiscuity than do domains in either bacteria or archaea. This implies that protein-protein interactions are greatest in eukaryotes, since there is a correlation between domain promiscuity (tendency to co-exist with other domains within proteins) and tendency to engage in domain-domain interactions between different proteins

[8, 13].

However, our results also show that the domains shared by three superkingdoms have the highet connectivity of all, suggesting that the functionalities common to all cellular organisms might be more involved in the protein-protein interactions. It is also a consequence that these domains have a long history so that they could develop their domain-domain networks more sophisticated than the domains shared by two superkingdoms and the domains unique to one superkingdom. However, domain connectivities among the 24 eukaryotic proteomes do not show the significant difference. It only showed that multi-cell organisms have slightly higher domain connectivities.

Capability to form versatile domain combinations

It was also investigated about the capability for a set of domains to form versatile domain combinations. We found that Archaea has higher capabilities than Bacteria and Eukaryota (both are about the same). If the domain definitions are sufficient enough in Archaea, then it proposes that Archaea might take this way to increase the capability of environmental adaption. Similarly, protists have higher capabilities and animals have lower capabilities, suggesting that those species with lower organismic complexity might develop a different approach to increase capability of environmental adaption.

The evolutionary history of domain rearrangement

Proteins can be regarded as a collection of domains, so the process of domain rearrangement could represent the protein evolution. This provides a way to interpret evolution in addition to the collection of mutations at individual residue site, because domains are the larger scale functional, structural and evolutionary units. The exponential relationship between the number of domains in a proteins and the number of proteins containing that number of domains (Figure 3-2) growth of domains in a protein suggests that the addition, deletion, and recombination of domains to a domain composition may follow some mechanism that can be described by a mathematical kinetic model.

Methods

Data source: UniProtKB v.2011_01

The domain content and taxonomical information of proteins used in this research were derived from Universal Protein Resource Knowledgabase (UniProtKB) v.2011_1 (published in January 11, 2011, http://www.uniprot.org/) [20]. UniProtKB is a comprehensive repository of proteins, consisting of two databases: Swiss-Prot, which is manually annotated, and TrEMBL, which is automatically annotated. With cross-reference to many databases and integrated protein information, UniProtKB has become a standard data source for many fields of protein bioinformatics research. In v.2011_1 of UniProtKB, it contained 13,593,921 protein entries, of which 524,420 are form Swiss-Prot and 13,069,501 are from TrEMBL. The statistics of the UniProtKB databases

can be found on the website. For this research, the flat data files of UniProtKB/Swiss-Prot and UniProt/TrEMBL were downloaded from UniProt website for further processing.

A list of complete proteomes can also be obtained from UniProt website. To be included in this list, the genome of an organism must be completely sequenced and with good quality of proteome data or good gene prediction models. In the v.2011_1 of UniProtKB, 1048 prokaryotic genomes (963 from Bacteria, 85 from Archea) and 129 eukaryotic genomes were included in the list of complete proteomes. The data files of these proteomes were retrieved from the UniProtKB data files.

Because UniProtKB databases provide extensive external cross-references to many databases, InterPro, Pfam, and Gene3D domain content were retrieved from the UniProtKB data files. The domain content then was parsed for further processing.

Programming in Perl

All the programming scripts in this research were written in Practical Extraction and Retrieval Language (Perl). Perl is a high-level and interpreted programming language, with powerful text manipulation capability and many available modules. Perl is widely used in the bioinformatics research. Also, the objective-oriented Swissknife Perl modules from UniProt were used to facilitate the parse of UniProt flat datafile.

3-way and 7-way division of UniProt Databse

In order to analyses the domain abundance and domain connectivity in all three superkingdoms, we divided the database in two different ways (Figure 3-8). The first one (3-way) is to assign each protein to one of the three superkingdom by its existence in corresponding superkingdom. The second one (7-way) is to divide the whole database by the taxonomical existence of domain compositions into 7 categories, and assign proteins to one of the 7 categories by their corresponding domain compositions.

Domain connectivity

Domain connectivity is defined as the number of connections of a domain in the domain co-occurrence network. This network can be drawn based on a taxonomical lineage, like a superkingdom or a single species. Domain co-occurrence describes the co-existence of two domains within a protein in this taxonomical lineage. In Figure 3-9, which is part or a domain co-occurrence network based on Eukaryota, the 23 connections (i.e. the degree of connectivity is 23) of IPR001508 are drawn and the numbers on these connections are the numbers of proteins where co-occurrences happen. IPR001508 is an InterPro domain of eukaryotic ionotropic glutamate receptor ion channels. Those domains with high co-occurrence numbers (more than 40) of connections with IPR001508 are also the members of eukaryotic ionotropic glutamate receptor ion channels. Most of the co-occurrences are found in Metazoa, only 18 co-occurrences (in 5 proteins) are not in Metazoa.

Selected eukaryotic proteomes

We extended our research from superkingdom level to some eukaryotic proteomics. Twenty-four eukaryotic species, including *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Bos Taurus*, *Gallus gallu*s, *Danio rerio*, *Branchiostoma floridae*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Nematostella vectensis*, *Monosiga brevicollis*, *Saccharomyces cerevisia*e, *Aspergillus oryzae*, *Schizosaccharomyces pombe*, *Ustilago maydis*, *Oryza sativa subsp. japonica*, *Arabidopsis thaliana*, *Physcomitrella patens subsp. patens*, *Ostreococcus tauri*, *Dictyostelium discoideum*, *Plasmodium chabaudi*, *Trypanosoma cruzi*, *Giardia lamblia ATCC 50803*, and *Trichomonas vaginalis*, were selected. They can be grouped into 4 major eukaryotic branches (metazoa, fungi, plants and protists). Their taxonomical relationship is illustrated in Figure 3-4. This selection, based on completely sequenced genomes, covers eukaryotes widely from single-cell organisms to multi-cell organism.

Computation of domain combination capability

This method is intended to provide an indication as for how capable it is of a finite set of domains to form versatile domain combinations. Given within a taxonomical unit, if there are N distinct domains and the maximal number of domains within a protein is K, the maximal number of domain combinations is:

$$\binom{N}{1} + \binom{N}{2} + \binom{N}{3} + ? A ? + \binom{N}{K}$$

Supposed there are M domain compositions in this taxonomical unit, the domain combination capability is:

$$\frac{M}{\binom{N}{1}+\binom{N}{2}+\binom{N}{3}+\cdots+\binom{N}{K}}$$

Practically, the domain combination capability is a very small number, since the maximal number of domain combination is a very large number. However, it is possible to compare the capabilities among different taxonomical units by simply comparing the relative magnitudes of these numbers. In this research, we only investigated the situation for InterPro domain definition. In UniProt database (v. 2011_01), there are 31 InterPro domains in a protein at the most. The maximal number of domain combinations would be the combinations of 1 domain to 31 domains in a protein. However, those domain compositions with large number of domains only account for a tiny fraction of total domain compositions. In order to avoid the bias of these domain compositions, we performed other 3 computations: combinations of 1 domain to14 domains in a protein (cover more than 99.9% of proteins in UniProt databse), combinations of 1 domain to 9 domains in a protein (cover more than 99% of proteins in UniProt databse), and combinations of 1 domain to 5 domains in a protein (cover more than 90% of proteins in UniProt databse). They are presented in the Results.

References

1. Song N, Sedgewick RD, Durand D: Domain architecture comparison for multidomain homology identification. *J Comput Biol* 2007, 14:496-516.

2. Bjorklund AK, Ekman D, Light S, Frey-Skott J, Elofsson A: Domain

rearrangements in protein evolution. *J Mol Biol* 2005, 353:911-923.

3. Moore AD, Bjorklund AK, Ekman D, Bornberg-Bauer E, Elofsson A: Arrangements in the modular evolution of proteins. *Trends Biochem Sci* 2008, 33:444-451.

4. Vogel C, Teichmann SA, Pereira-Leal J: The relationship between domain duplication and recombination. *J Mol Biol* 2005, 346:355-365.

5. Ger MF, Rendon G, Tilson JL, Jakobsson E: Domain-based identification and analysis of glutamate receptor ion channels and their relatives in prokaryotes. *PLoS One* 2010, 5:e12827.

6. Rendon G, Ger M-F, Kantorovitz R, Natarajan S, Tilson J, Jakobsson E: Understanding the "Horizontal Dimension" of Molecular Evolution to Annotate, Classify, and Discover Proteins with Functional Domains. *JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY* 2010, 25:82-94.

7. Lin K, Zhu L, Zhang DY: An initial strategy for comparing proteins at the domain architecture level. *Bioinformatics* 2006, 22:2081-2086.

8. Basu MK, Carmel L, Rogozin IB, Koonin EV: Evolution of protein domain promiscuity in eukaryotes. *Genome Res* 2008, 18:449-461.

9. Lee B, Lee D: Protein comparison at the domain architecture level. *BMC Bioinformatics* 2009, 10 Suppl 15:S5.

10. Vogel C, Chothia C: Protein family expansions and biological complexity. *PLoS Comput Biol* 2006, 2:e48.

11. Marsden RL, Lee D, Maibaum M, Yeats C, Orengo CA: Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. *Nucleic Acids Res* 2006, 34:1066-1080.

12. Ekman D, Bjorklund AK, Frey-Skott J, Elofsson A: Multi-domain proteins in

the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol* 2005, 348:231-243.

13. Basu MK, Poliakov E, Rogozin IB: Domain mobility in proteins: functional and evolutionary implications. *Brief Bioinform* 2009, 10:205-216.

14. Weiner J, 3rd, Moore AD, Bornberg-Bauer E: Just how versatile are domains? *BMC Evol Biol* 2008, 8:285.

15. Wuchty S: Scale-free behavior in protein domain networks. *Mol Biol Evol* 2001, 18:1694-1702.

16. McDowall J, Hunter S: InterPro protein classification. *Methods Mol Biol* 2011, 694:37-47.

17. Lees J, Yeats C, Redfern O, Clegg A, Orengo C: Gene3D: merging structure and function for a Thousand genomes. *Nucleic Acids Res* 2010, 38:D296-300.

18. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A: The Pfam protein families database. *Nucleic Acids Res* 2008, 36:D281-288.

19. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al: InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009, 37:D211-215.

20. Consortium TU: The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 2010, 38:D142-148.

21. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: The large-scale organization of metabolic networks. *Nature* 2000, 407:651-654.

22. Tordai H, Nagy A, Farkas K, Banyai L, Patthy L: Modules, multidomain proteins and organismic complexity. *FEBS J* 2005, 272:5064-5078.

23. Adami C, Ofria C, Collier TC: Evolution of biological complexity. *Proc Natl Acad Sci U S A* 2000, 97:4463-4468.

Tables

Table 3-1. - Domain content in UniProtKB Database

InterPro, Pfam, and Gene3D domains were retrieved from UniProtKB 2011_01. For each domain definition, the total number of domains and domain compositions were parsed. The average degrees of connectivity in the domain-domain co-occurrence network were calculated. The average numbers of domains for proteins carrying corresponding domain definition were calculated, excluding proteins without domain definition and counting repeats for Pfam and Gene3D defitions. Standard deviations for each average number are listed in the parentheses. See Methods for the further details

UniProt 2011_01 (13,593,921 proteins)		
Proteins carrying InterPro domain 10,535,894 proteins (77.50%)	Proteins carrying Pfam domain 9,977,578 (73.40%)	Proteins carrying Gene3D domain 4,479,982 (32.96%)
21,091 domains	11,464 domains	1,147 domains
degree of connectivity/domain: 23.10 (standard deviation: 79.00)	degree of connectivity/domain : 7.71 (standard deviation: 21.91)	degree of connectivity/domain: 9.92 (standard deviation 21.78)
153,165 domain compositions	50,369 domain compositions 77,561 domain compositions(if counting repeats)	6,869 domain compositions 14,451 domain compositions(if counting repeats)
2.77 domains/protein (standard deviation 2.08)	1.48 domains/protein (standard deviation 1.25)	1.47 domain/protein (standard deviation 1.19)

Table 3-2. – Average of numbers of InterPro domains and average of degrees of connectivity

The averages of numbers of InterPro domains per protein and their standard deviations were calculated within 24 eukaryoitc genomes. So were the averages of degrees of connectivity per InterPro domain. Then, these 24 were grouped into 4 groups(animals, fungi, plants and protists). The averages were calculated from the corresponding groups.

Species	Taxid	Average of numbers of domains per protein	Standard deviation	Average of degrees of connectivity per domain	Standard deviation
Homo sapiens	9606	3.33	0.29	8.14	16.9
Mus musculus	10090	3.44	0.40	7.91	16.7
Rattus norvegicus	10116	3.51	0.36	7.66	15.2
Bos Taurus	9913	3.36	0.36	7.09	13.7
Gallus gallus	9031	3.62	0.44	7.58	13.4
Danio rerio	7955	3.45	0.44	7.32	14.1
Branchiostoma floridae	7739	3.32	0.43	9.83	21.5
Drosophila melanogaster	7227	3.10	0.50	6.90	11.7
Caenorhabditis elegans	6239	2.71	0.41	6.61	10.9
Nematostella vectensis	45351	2.85	0.34	6.67	11.7
Monosiga brevicollis	81824	3.13	0.49	7.04	13.4
Saccharomyces cerevisiae	4932	2.91	0.60	4.71	6.10
Aspergillus oryzae	5062	2.75	0.59	5.18	7.43
Schizosaccharomyces pombe	4896	2.95	0.51	4.86	6.74
Ustilago maydis	5270	2.89	0.53	5.17	7.33
Oryza sativa	39947	2.72	0.16	6.76	12.6
Arabidopsis thaliana	3702	2.69	0.22	5.71	9.87
Physcomitrella patens	145481	2.73	0.34	5.17	8.22
Ostreococcus tauri	70448	2.87	0.45	6.05	9.54
Dictyostelium discoideum	44689	2.89	0.70	6.03	11.1
Plasmodium chabaudi	5825	2.13	0.40	3.47	3.62
Trypanosoma cruzi	5693	2.55	0.33	4.46	5.75
Giardia lamblia	184922	2.72	0.04	3.82	3.28
Trichomonas vaginalis	5722	2.42	0.20	4.50	6.38
Animals		3.27	0.29	7.57	0.94
Fungi		2.87	0.09	4.98	0.24
Plants		2.75	0.08	5.92	0.67
Protists		2.64	0.36	4.95	1.49

Table 3-3. – The distributions of numbers of InterPro domains and the distributions of degrees of connectivity

The distributions of numbers of InterPro domains per protein and degrees of connectivity per InterPro domain were estimated by exponential law distribution and power law distributions, respectively. The parameters of respective distributions and fitness are listed in this table. Then, these 24 were grouped into 4 groups(animals, fungi, plants and protists). The averages of the parameters were calculated from the corresponding groups.

	Distribution of numbers of domainsper protein			Distribution of degrees of connectivity per domain		
	Equation		Fitness			Fitness
Homo sapiens	$y = 36963e^{-0.411x}$	-0.411	$R^2 = 0.9838$	$y = 2727x^{-1.555}$	-1.555	$R^2 = 0.8600$
Mus musculus	$y = 24518e^{-0.391x}$	-0.391	$R^2 = 0.9819$	$y = 2921.7x^{-1.568}$	-1.568	$R^2 = 0.8938$
Rattus norvegicus	$y = 12381e^{-0.389x}$	-0.389	$R^2 = 0.9788$	$y = 2805.5x^{-1.565}$	-1.565	$R^2 = 0.8817$
Bos Taurus	$y = 9170.5e^{-0.416x}$	-0.416	$R^2 = 0.971$	$y = 2869x^{-1.607}$	-1.607	$R^2 = 0.8808$
Gallus gallus	$y = 4610.8e^{-0.346x}$	-0.346	$R^2 = 0.9662$	$y = 2555.2x^{-1.596}$	-1.596	$R^2 = 0.8800$
Danio rerio	$y = 10860e^{-0.385x}$	-0.385	$R^2 = 0.9703$	$y = 2591.4x^{-1.588}$	-1.588	$R^2 = 0.8744$
Branchiostoma floridae	$y = 8001.9e^{-0.359x}$	-0.359	$R^2 = 0.9799$	$y = 1659.3x^{-1.439}$	-1.439	$R^2 = 0.8695$
Drosophila melanogaster	$y = 10637e^{-0.391x}$	-0.391	$R^2 = 0.9636$	$y = 2568.1x^{-1.630}$	-1.630	$R^2 = 0.8829$
Caenorhabditis elegans	$y = 7943.8e^{-0.432x}$	-0.432	$R^2 = 0.9858$	$y = 2454.1x^{-1.652}$	-1.652	$R^2 = 0.8903$
Nematostella vectensis	$y = 8809.6e^{-0.457x}$	-0.457	$R^2 = 0.9603$	$y = 2627.3x^{-1.635}$	-1.635	$R^2 = 0.8958$
Monosiga brevicollis	$y = 3171e^{-0.401x}$	-0.401	$R^2 = 0.9673$	$y = 1578.9x^{-1.531}$	-1.531	$R^2 = 0.8657$
Saccharomyces cerevisiae	$y = 2444.1e^{-0.431x}$	-0.431	$R^2 = 0.9237$	$y = 2401.8x^{-1.776}$	-1.776	$R^2 = 0.8813$
Aspergillus oryzae	$y = 2117e^{-0.342x}$	-0.342	$R^2 = 0.9078$	$y = 2695.8x^{-1.763}$	-1.763	$R^2 = 0.9171$
Schizosaccharomyces pombe	$y = 1732e^{-0.415x}$	-0.415	$R^2 = 0.9106$	$y = 2331.6x^{-1.765}$	-1.765	$R^2 = 0.8887$
Ustilago maydis	$y = 2550.2e^{-0.440x}$	-0.440	$R^2 = 0.9700$	$y = 2174.9x^{-1.720}$	-1.720	$R^2 = 0.8746$
Oryza sativa	$y = 66939e^{-0.600x}$	-0.600	$R^2 = 0.9655$	$y = 1953.7x^{-1.573}$	-1.573	$R^2 = 0.8540$
Arabidopsis thaliana	$y = 24794e^{-0.513x}$	-0.513	$R^2 = 0.9789$	$y = 2418.6x^{-1.697}$	-1.697	$R^2 = 0.8715$
Physcomitrella patens	$y = 16083e^{-0.550x}$	-0.550	$R^2 = 0.9713$	$y = 2693.3x^{-1.752}$	-1.752	$R^2 = 0.8813$
Ostreococcus tauri	$y = 3360.6e^{-0.471x}$	-0.471	$R^2 = 0.9758$	$y = 1955.7x^{-1.657}$	-1.657	$R^2 = 0.8876$
Dictyostelium discoideum	$y = 2687.1e^{-0.364x}$	-0.364	$R^2 = 0.9012$	$y = 1782.5x^{-1.617}$	-1.617	$R^2 = 0.8692$
Plasmodium chabaudi	$y = 3378.3e^{-0.593x}$	-0.593	$R^2 = 0.9248$	$y = 1693.5x^{-1.937}$	-1.937	$R^2 = 0.9138$
Trypanosoma cruzi	$y = 8541.9e^{-0.553x}$	-0.553	$R^2 = 0.9561$	$y = 1736.2x^{-1.783}$	-1.783	$R^2 = 0.8957$
Giardia lamblia	$y = 3676e^{-0.636x}$	-0.636	$R^2 = 0.9704$	$y = 1312.2x^{-1.928}$	-1.928	$R^2 = 0.8522$
Trichomonas vaginalis	$y = 25690e^{-0.675x}$	-0.675	$R^2 = 0.9751$	$y = 1357.6x^{-1.723}$	-1.723	$R^2 = 0.8813$
		Mean	Standard Deviation		Mean	Standard Deviation
Animals		-0.3977	0.0328		-1.5835	0.0601
Fungi		-0.4070	0.0446		-1.7560	0.0247
Plants		-0.5335	0.0548		-1.6698	0.0753
Protists		-0.5370	0.1270		-1.7532	0.1637

Figures

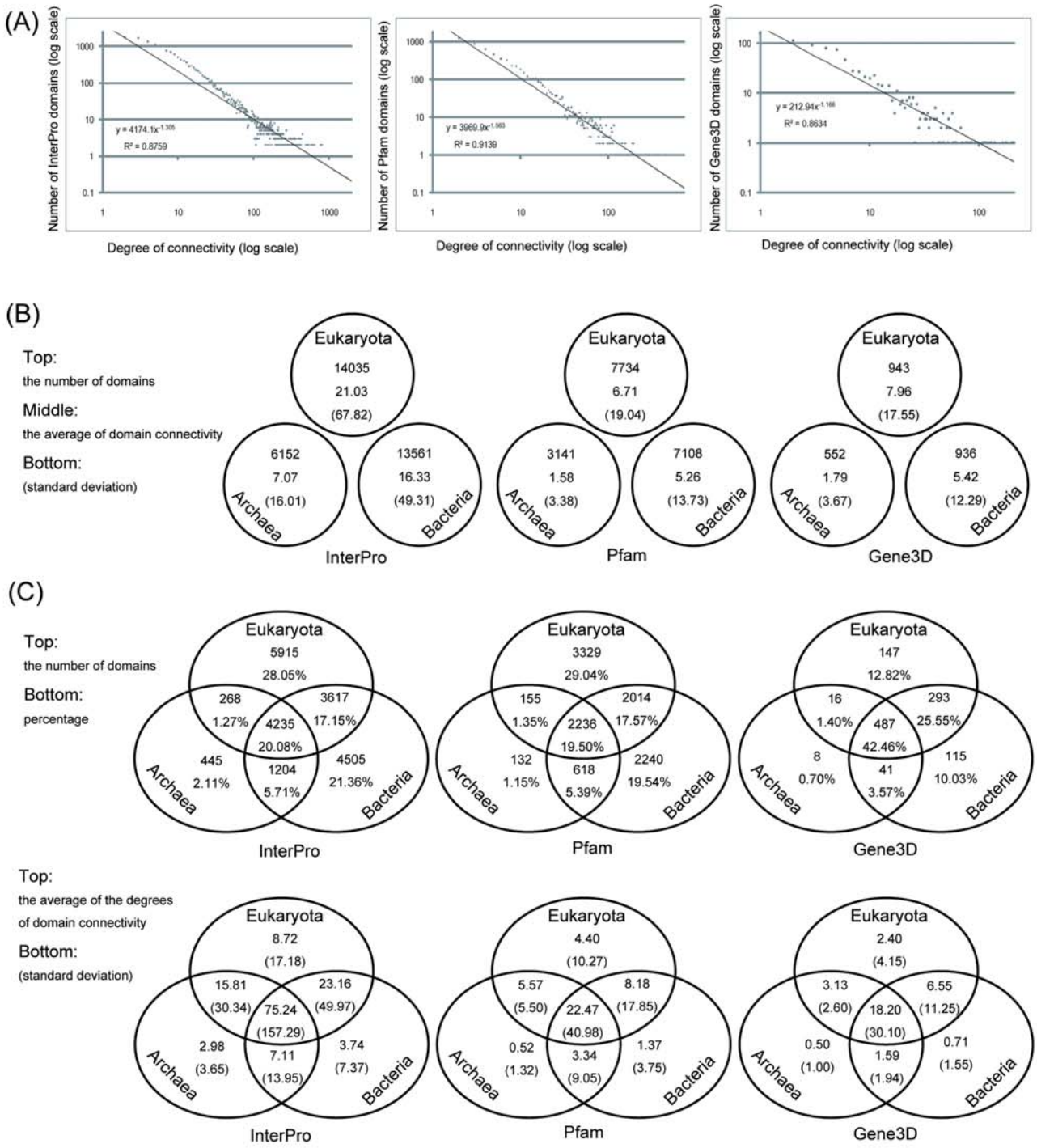


Figure 3-1. - The statistics of InterPro, Pfam and Gene3D domains in the domain-domain co-occurrence networks
(A) The distributions of the degrees of connectivity in domain-domain co-occurrence networks are showed (left: InterPro domains, middle: Pfam domains, right: Gene3D domains). In each graph, best fit of line is drawn with the equation and R square value. All these three fitting lines follow the power law, which showing the characteristic of scale-free network. (B) The top three Venn diagrams are the taxonomical distributions of domains (left: InterPro domains, middle: Pfam domains, right: Gene3D domains). Domains were categorized according to existences in one, two, or three superkingdoms. The percentages to the total domains are listed in the parentheses. The bottom Venn diagrams are the average degrees of connectivity in each category. Outside the Venn diagrams are the average degrees of connectivity in the each lineage, which is an integrated number of four categories. Standard deviations for each average number are listed in the parentheses

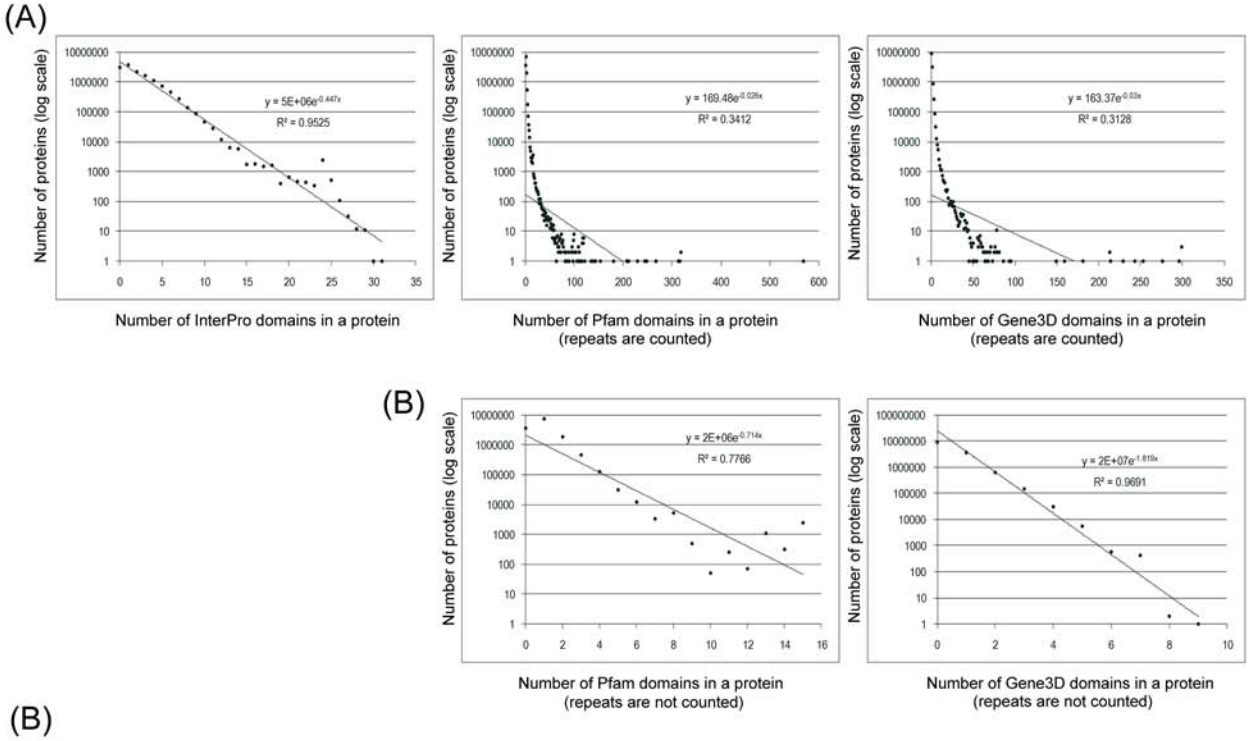


Figure 3-2. - The distributions of the number of domains within a protein
The distributions of the number of domains within a protein are graphed (left: InterPro domains, middle: Pfam domains, right: Gene3D domains). For the Pfam and Gene3D domains, repetitive domains are counted in the domain compositions. (B) The distributions of the number of domains within a protein are graphed (Pfam domains, right: Gene3D domains). Repetitive domains are treated as one domain in the domain compositions. In each graph, best fit of line is drawn with the equation and R square value. All these three fitting lines follow the exponential law.

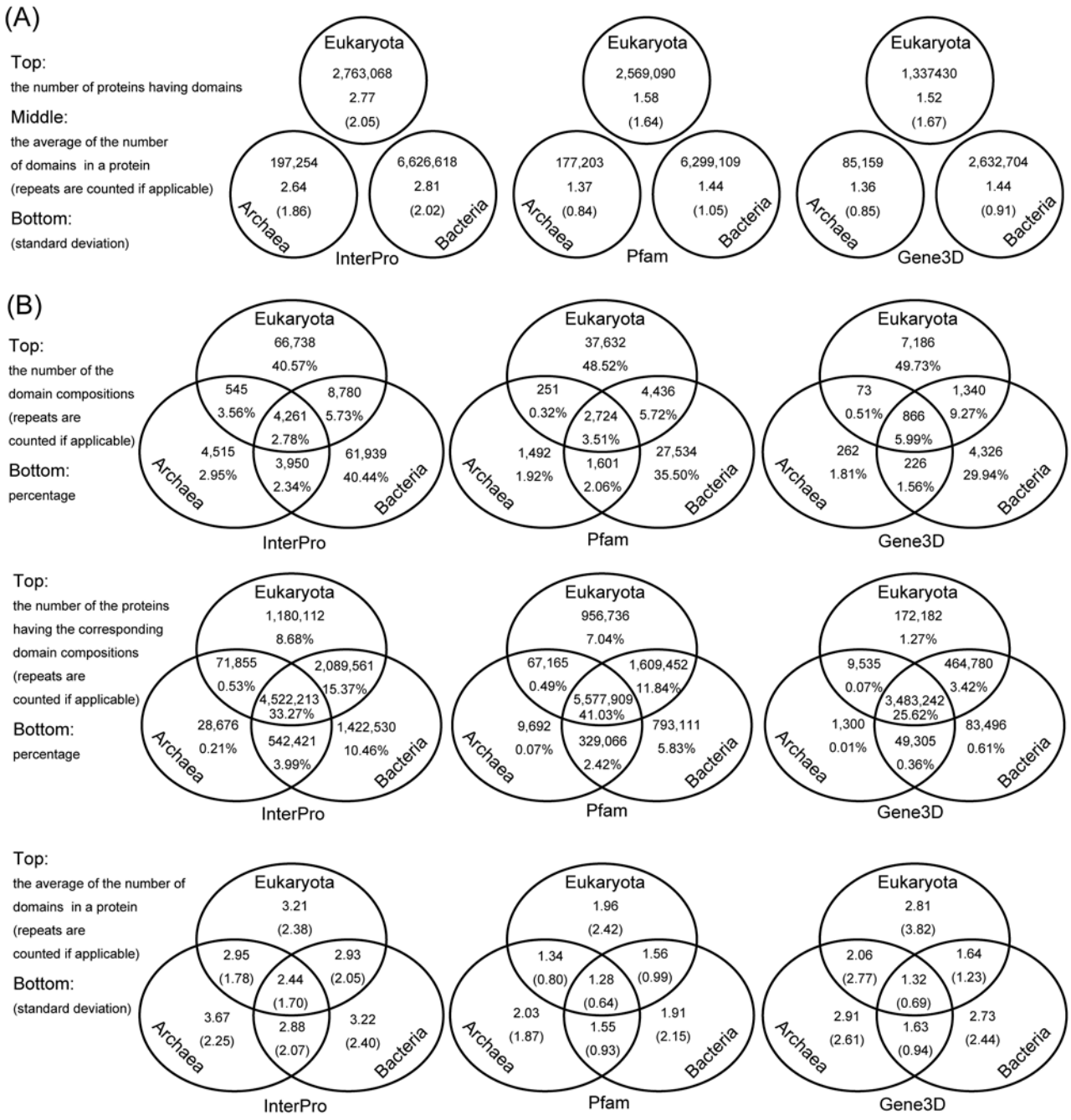


Figure 3-3. - The statistics of InterPro, Pfam and Gene3D domain compositions
The first row is Venn diagrams of the distributions of domain compositions (left: InterPro domains, middle: Pfam domains, right: Gene3D domains). The percentage of each category is also listed. The second row is the Venn diagrams of the number of the proteins carrying the respective domain compositions. The percentage of each category is also listed. The third row is the Venn diagrams of the average of domains per protein in each category. Standard deviations for each average number are listed in the parentheses.

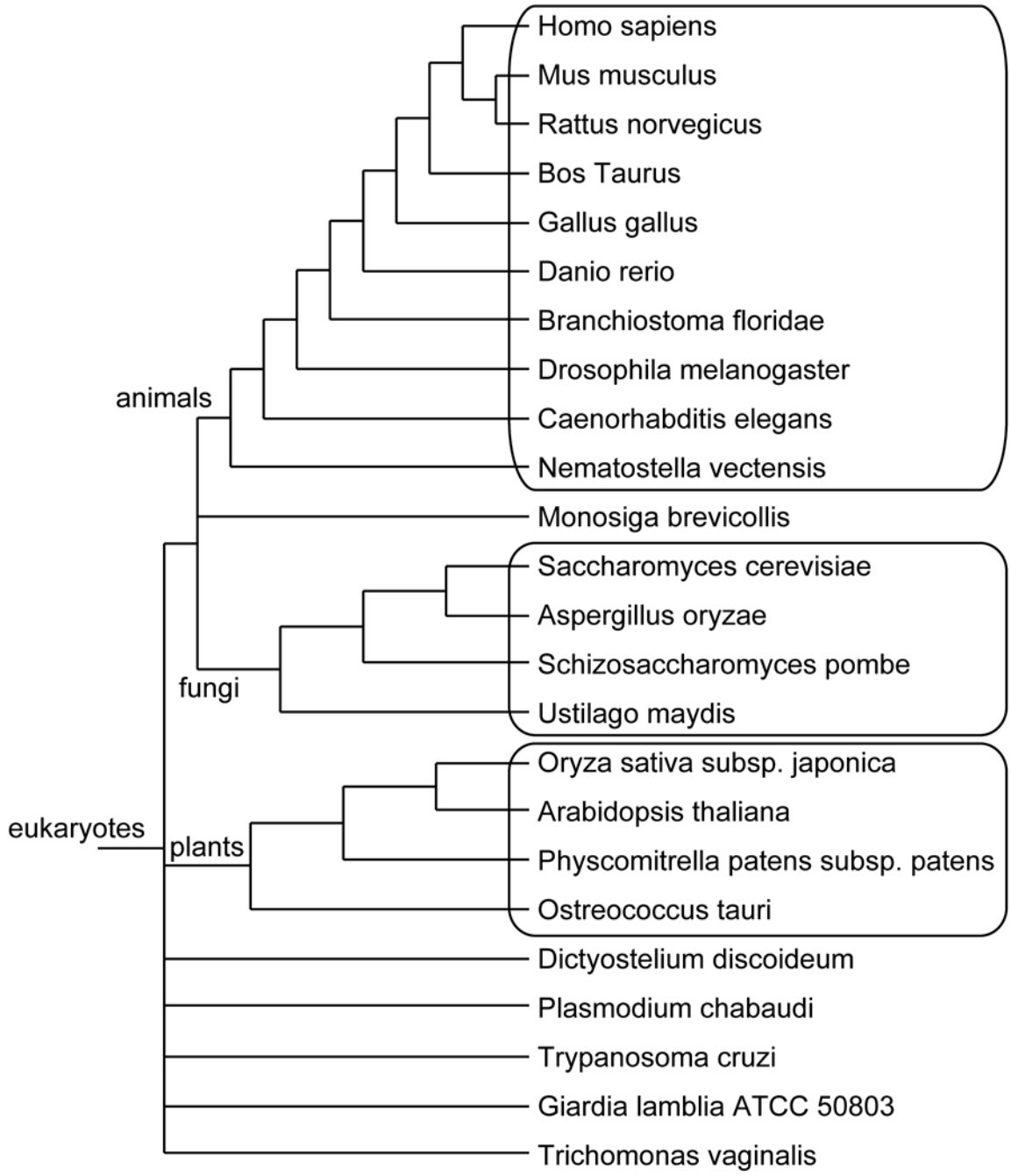


Figure 3-4. - The taxonomical hierarchy of 24 eukaryotic species
This taxonomical hierarchy was derived from National center for Biological Information website. Out of 24, 10 are animals; 4 are fungi; 4 are plants; and the other 6 are protists.

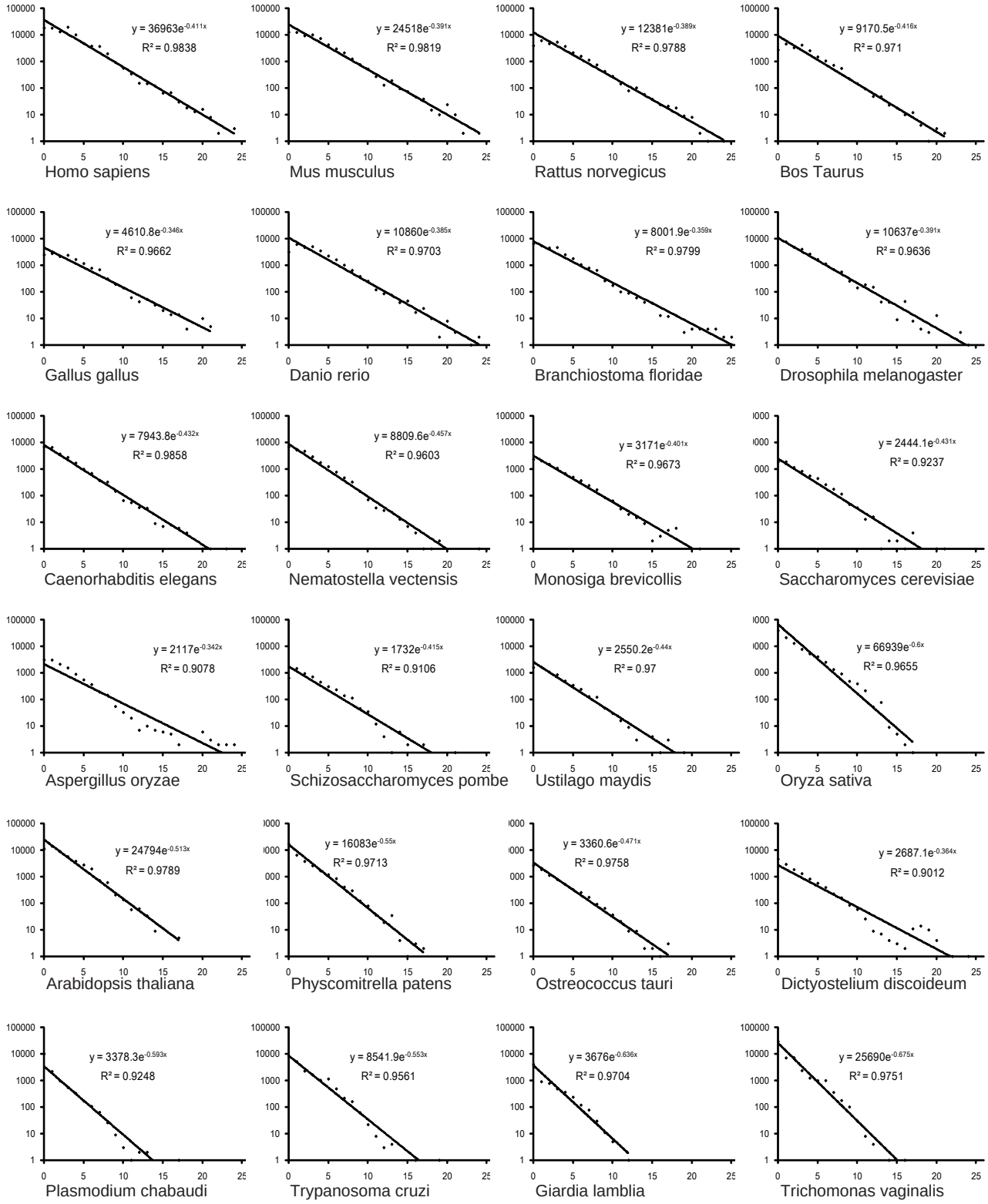


Figure 3-5. - The distributions of numbers of InterPro domains in 24 eukaryotic proteomes

The figures of the expotential-law distributions of 24 eukaryotic proteomes are listed here. The parameters and the fitness of the distributions are also included in the figures.

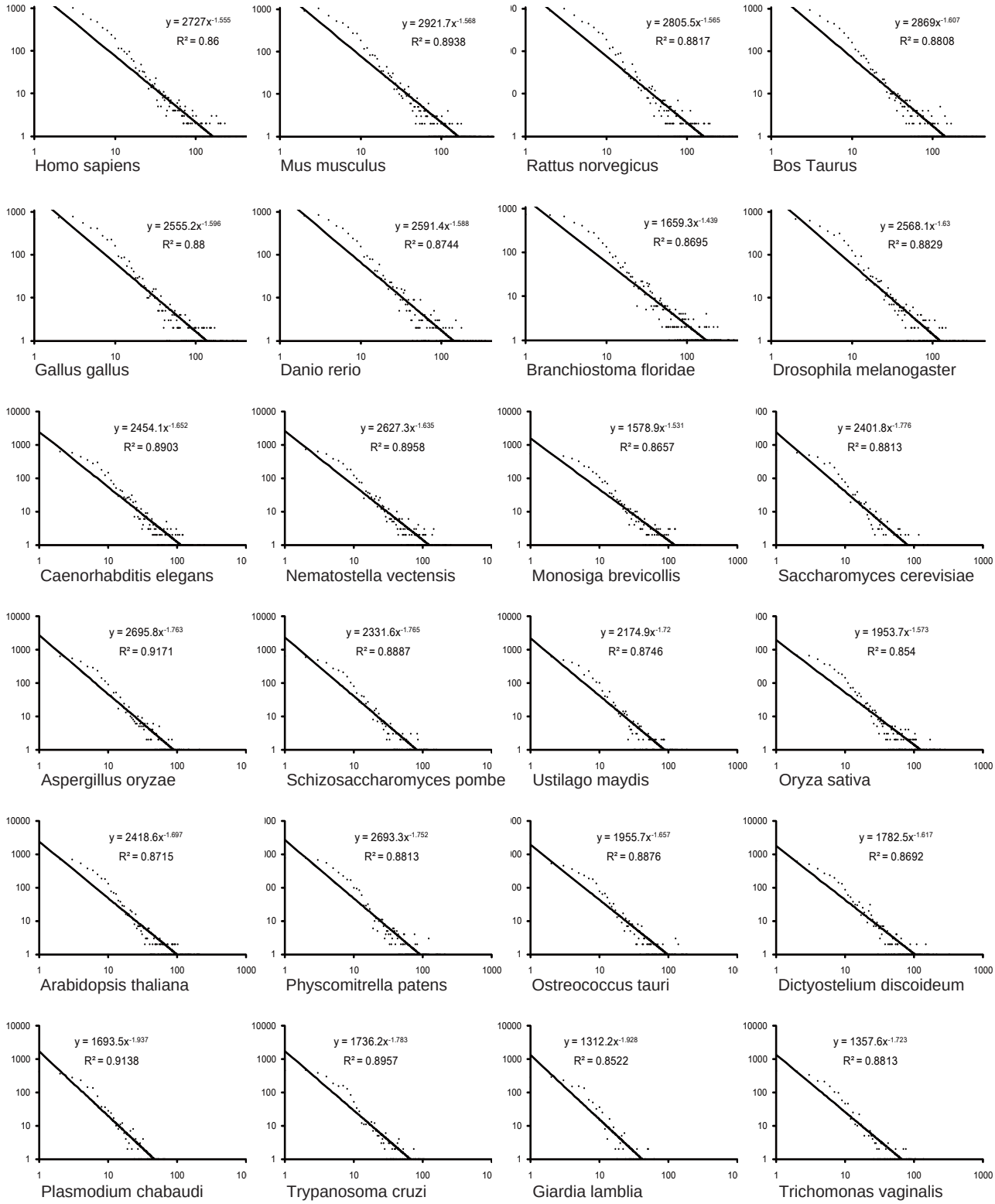
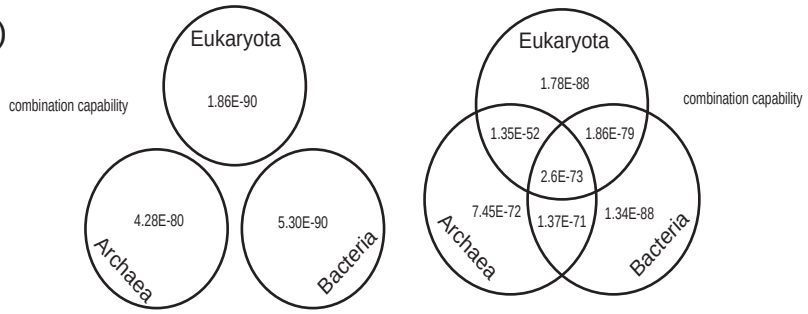


Figure 3-6. - The distributions of degrees of connectivity of InterPro domains in 24 eukaryotic proteomes
The figures of the power-law distributions of 24 eukaryotic proteomes are listed here. The parameters and the fitness of the distributions are also included in the figures.

(A)



(B)

	Tax ID	Combination Capability	Taxnomy	Mean	Standard Deviation
Homo sapiens	9606	3.69E-84			
Mus musculus	10090	3.80E-84			
Rattus norvegicus	10116	7.47E-84			
Bos Taurus	9913	2.98E-83			
Gallus gallus	9031	8.57E-81	Animals	6.41E-80	1.69E-79
Danio rerio	7955	2.86E-82			
Branchiostoma floridae	7739	1.44E-80			
Drosophila melanogaster	7227	5.86E-80			
Caenorhabditis elegans	6239	5.43E-79			
Nematostella vectensis	45351	1.61E-80			
Saccharomyces cerevisiae	4932	1.24E-76			
Aspergillus oryzae	5062	1.59E-77	Fungi	1.56E-76	1.1E-76
Schizosaccharomyces pombe	4896	2.64E-76			
Ustilago maydis	5270	2.20E-76			
Oryza sativa	39947	1.13E-79			
Arabidopsis thaliana	3702	1.15E-79	Plants	2.72E-76	5.44E-76
Physcomitrella patens	145481	1.12E-79			
Ostreococcus tauri	70448	1.09E-75			
Monosiga brevicollis	81824	1.35E-76			
Dictyostelium discoideum	44689	7.89E-77			
Plasmodium chabaudi	5825	8.86E-69	Protists	6.85E-66	1.68E-65
Trypanosoma cruzi	5693	3.12E-72			
Giardia lamblia	184922	4.11E-65			
Trichomonas vaginalis	5722	9.87E-71			

Figure 3-7. - The capability of forming versatile InterPro domain combinations

(A) The numbers representing the capability of forming versatile InterPro domain combinations in the 7-way divisions of all domain compositions and in the 3-way divisions of all domain compositions. (B) The numbers representing the capability of forming versatile InterPro domain combinations in 24 eukaryotic proteomes and the averages of 4 groups (animals, fungi, plants, and protists).

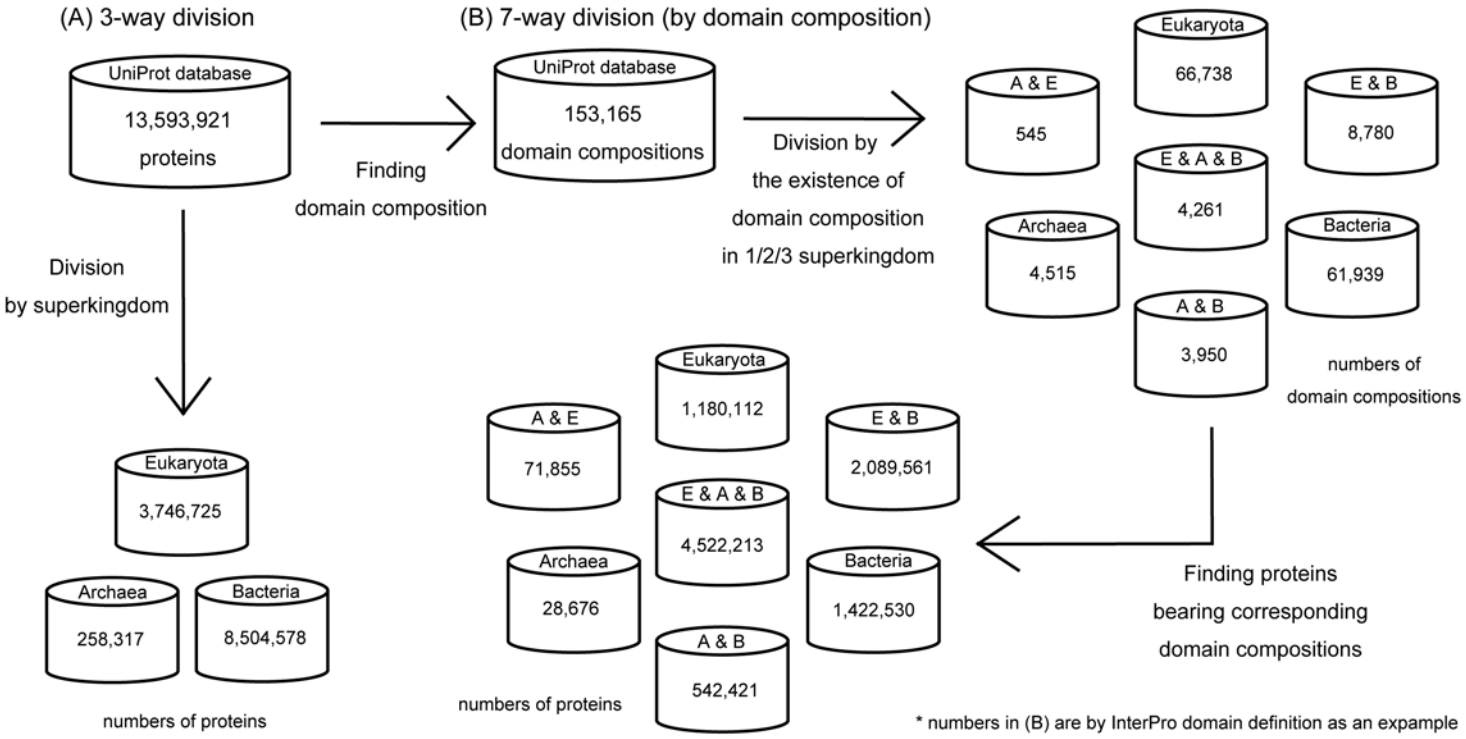


Figure 3-8. - 3-way and 7-way divisions of the UniProt Database
This figure shows the methodology of dividing the UniProt databse by the taxonomy (i.e. the superkingdom) into 3 smaller databases or 7 smaller database.

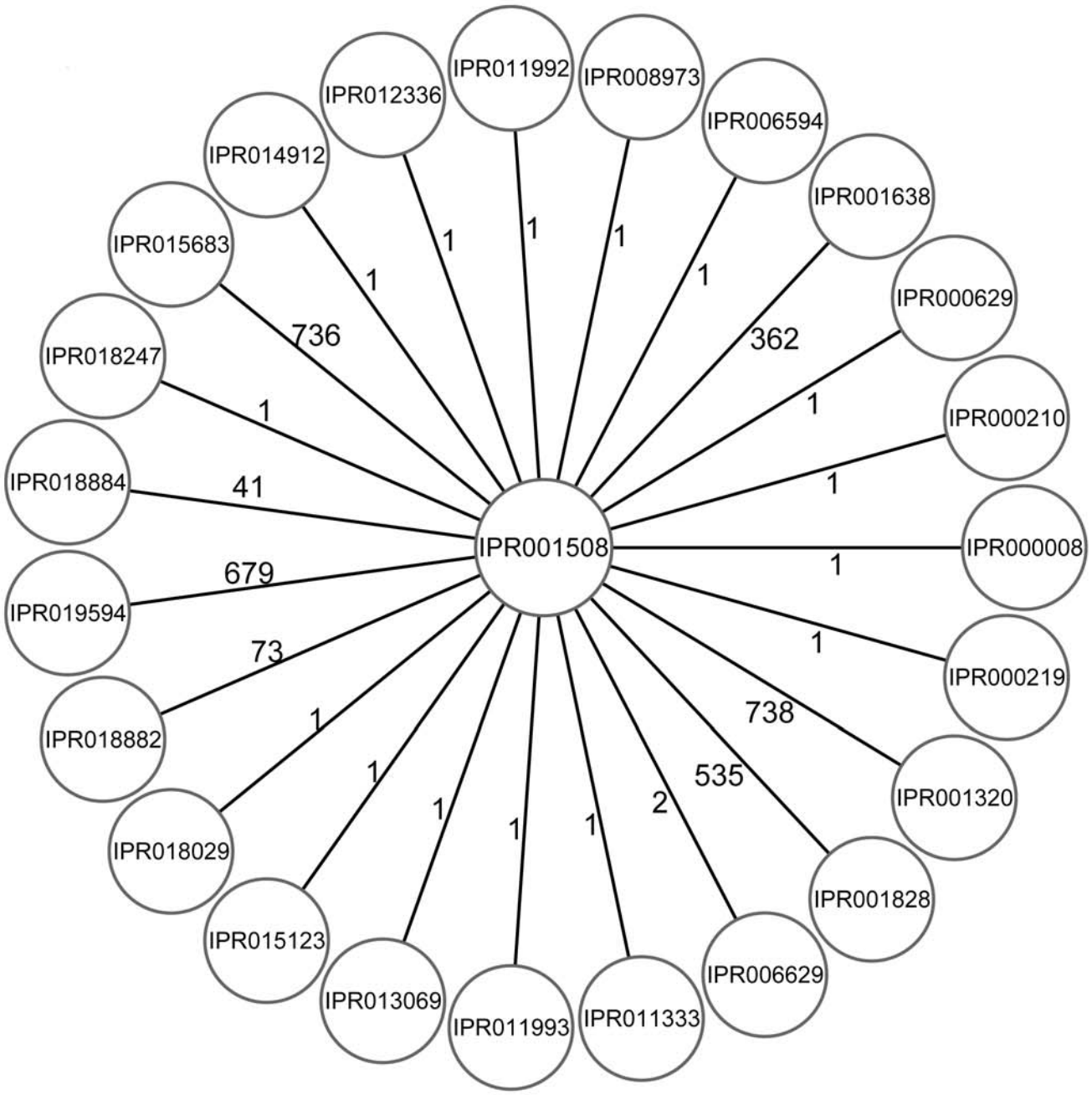


Figure 3-9. - An example of domain connectivity
Domain co-occurrence is the co-existence of two domains within a protein. In this network, nodes are domains and edges are the co-occurrences. This diagram is part of a domain co-occurrence network. The central node is an InterPro domain IPR001508: NMDA receptor, which is commonly found in eukaryotic ionotropic glutamate receptor ion channels. The degree of connectivity of this domain is 23, which means it has 23 edges (connections, co-occurence). The numbers on edges represent the number of the proteins this co-occurrence happen. The nodes on the circumference may have varying numbers of degrees of connectivity which are not shown on this diagram. All of these 3181 co-occurrences happen in 741 proteins. The majority of these proteins are in a narrower taxonomical lineage, Metazoa. However, 5 proteins are the exceptions.

CHAPTER 4:
The comparison of proteins by domain content and its application

Abstract

In this chapter we explore the use of domain content as a measure of evolutionary distance. The view of evolution underlying this chapter is that domain fission, fusion, and replications are important components of the evolutionary process. Comparing domain composition is an alignment-free method of computing evolutionary distance. The concept of "cosine similarity" treats each domain combination as a vector, and evolutionary distance as the distance in the vector space between combinations. I analyse trends of cosine similarity across the entire UniProt database and explore relationship between cosine similarity and OMA orthology. Because of the correspondence between cosine similarity and orthology in those proteins included in the OMA orthology databases, I suggest that "cosine identity (similarity=1)" may be a reasonable surrogate for orthology in protein space not covered by OMA. These concepts are applied to families of ion channels, as an example.

Background

Due to a flood of protein sequences produced by numerous genome projects, function annotation of new proteins and evolutionary relationship between protein families are imperious to modern biology. To date, more than 18 million protein sequences are

available in public database [1]. Experimental methods could only satisfy a small portion of such needs, so bioinformatics tools are needed to reliably identify functionally equivalent proteins and homologously related proteins. The more common search approach for this purpose is based on sequence similarity, e.g. with BLAST or Psi-BLAST. However, functional equivalence is not tied to a significant sequence similarity, especially for those proteins diverged apart early in evolution. And, to make identification more confusing, proteins may be identified with high sequence similarity, but have evolved into different functions. Such errors in annotations could spread through databases [2], and contaminate the further researches. Therefore, search based on similarity of domain content provides another prospect for correct identification of homologs.

Domain-based methods use information of the domains, which are the basic units of function, structure and evolution, to serve the similar works. The previous researches by domain-based methods have shown that comparing domain architectures is a useful approach to identify evolutionarily distant homologs [3], especially for muti-domain proteins. However, these approaches are challenged by promiscuous domains, which co-occur with other domains in many ways to expand the versatileness of proteins[4]. They are mainly for the auxiliary functions, but not related to homology [5]. Hence, the comparison of domain architecture should lower the importance of promiscuous domain.

Several methods have been developed to compare domain architecture [6-10]. Different strategies were applied to correctly mimic the process of evolution. It is believed that a gene is not the unit of orthology, but is a domain [11, 12]. Throught fussion, fission, recombination of domains, as well as the creation of new domains,

proteomes are expanded. This modularity allows the evolution of diverse functions through combinatorial rearrangement, so the tracking of the combination process could be done by the comparisons of a collection of domain architectures (taking the order of domains) or domain compositions (not taking the order of domains) in a pair-wise way, just like the first step of multiple sequences alignment. Therefore, in addition to compare two domain architectures/compositions visually, the distance should also be derived numerically as a foundation of rebuilding the history of domain rearrangement.

It is not clear yet about the mechanisms of domain rearrangements. The emergence of particular domain architectures/compositions may have shaped by the selective forces. From the previous research, it is suggested that the domain rearrangements happened in a random way [13]. The modelling of this process has not been done probably because of the difficulty of complicated interactions among domains and the uncertain mechanism mediating the domain rearrangement. This work will be a key issue in the research of domain analysis.

In this research, we first did an assignment of significance score to every domain by three different domain definitions (InterPro, Pfam and Gene3D). Two methods were proposed by taking promiscuity into consideration or not. Then, a method of comparing two domain compositions by measuring the cosine similarity was developed. We showed that the similarity scores were more useful in identifying the evolutionary relationship if we took promiscuity into consideration. Also, we made a case study of 16 human ion channels to show that the cosine similarity scores can be used to construct a network of protein families, which might help explain the evolutionary relationship among these proteins.

Results

Computation of inverse domain frequency scores, domain evolutionary significance scores for domains and cosine similarity scores for pairs of domain compositions

From the previous chapter, the domain frequency (the number of domain occurrences in a database) and the number of distinct domains for InterPro, Pfam and Gene3D domains have been parsed from UniProt Databases v.2011_01, as have the domain compositions. Here, we developed two ways to compute the scores which can represent the significance of domains. The first way is to take the inverse domain frequency (IDF) scores as an index which can show how informative a domain is in determining the function by domain definitions. The second way is to take the product of inverse domain frequency (IDF) and inverse number of distinct partners (promiscuity) as an index that serves the similar purpose for evolutionary closeness (named as Domain Evolutionary Significance score, DES score). In DES, the impact of promiscuity on the cosine similarity scores was calculated by introducing the number of distinct partner domains. Both methods are derived from the field of information retrieval to determine the significance of words in text content. See Methods for details.

To evaluate the relatedness of two domain compositions, we applied the cosine similarity scores, which is a measure of similarity between two vectors (a vector is a set of domains). The score of each domain in a vector was taken from either the IDF scores or the DES scores from the above. Although it is believed that InterPro

domain definition covers a broader space of a database and a more comprehensive functionality, we still computed the cosine similarity scores by InterPro, Pfam, and Gene3D domain definitions (repetitious domains are counted if applicable). From Table 1, there are 195,916,056 pairs of InterPro domain compositions sharing at least one domain in common. The corresponding numbers for Pfam and Gene3D domain compositions are 19,541,013 and 3,957,161, respectively.

As for InterPro result, in cosine similarity based on IDF scores, the maximal similarity score is 0.9945 and the minimal similarity score is 0.0102. In cosine similarity scores based on DES scores, the maximal similarity score is 0.999999988 and the minimal similarity score is 2.3673E-8. This suggests that the cosine similarity scores based on DES scores create a higher resolution than on IDF scores. The results on Pfam and Gene3D domain definition suggest the same inference.

The average of InterPro cosine similarity scores on DES is much lower than that on IDF (0.2046 to 0.0370), indicating that domains with high promiscuity (i.e. large number of distinct partner domains in the same protein) affect the comparison of domain compositions in a significant way. Once this effect is removed, only pairs of domain compositions sharing less-promiscuous domains could have high scores of cosine similarity. As for Pfam and Gene3D, the decreases of averages from IDF to DES are not as much as that of InterPro (0.2766 to 0.1159, and 0.2611 to 0.1256, respectively), reflecting the fact that InterPro definition is more redundant, so that there are more highly promiscuous domains by this domain definition.

From Figure 1, the distributions of these cosine similarity scores were investigated. As for InterPro domain definition, the distribution of scores on IDF follows an exponential law (fitness of 0.81) and the peak is around 0.8. However, the

distribution of scores on DES shows a power law trend (fitness of 0.97). This reflects the difference of the characteristics between IDF and DES. As for Pfam and Gene3D domain definitions, the same phenomenon was observed.

The relatedness between cosine similarity scores by InterPro and Orthologous Matrix (OMA)

In this section is examined the degree of agreement between cosine similarity score and the orthology suggested by Othology MAtrix Project (OMA). We parsed the OMA information from the UniProt database v.2011_01. From Table 2, the number of OMA groups (388,214) is larger than the number of InterPro domain compositions (153,165); even larger than the number of Pfam domain compositions (77,561); and even much larger than the number of Gene3D domain compositions (14,451). A single domain composition group may contain more than one OMA orthology group. Taking an example from InterPro data, there are 22,087 InterPro domain composition group containing more than one OMA groups. Cosine similarity between domain composition groups may be used in determining the evolutionary relationship between the corresponding OMA groups. It is noted that most of the domain compositions are without OMA for the reason that OMA groups are constructed by about 1100 species, which is from a smaller space than UniProt. In the larger space, domain composition may be used to construct reasonably good orthology groups among samples that are not included in OMA.

The majority of the OMA groups are with exactly one domain composition (e.g. 57.66% on InterPro domain definition). As for the OMA groups containing more than one InterPro domain compositions, the averages of cosine similarity scores are significantly higher than the average of the database, suggesting the agreement

between OMA and cosine similarity. Moreover, the difference of averages on DES is larger than the difference of averages on IDF (0.5771/0.0370 to 0.7348/0.2406). This makes the scores on DES is more suitable in determining the evolutionary relationship than the scores on IDF. This holds for all three domain definitions. Because Gene3D covers a smaller space of the database, it is less useful for comprehensive evolutionary studies.

A case study: a network of 16 ion channels

We applied the method of cosine similarity on the 16 ion channel families in the database of the International Union of Basic and Clinical Pharmacology (IUPHAR, http://www.iuphar-db.org/). Figure 2 shows the network of cosine similarity scores (based DES scores) for the domain combinations of the human versions of 16 protein families. Eight out of ten P-loop ion channels can be built into a network showing the cosine similarity scores. Two types of P-loop ion channels (inward rectifiers and ionotropic glutamate receptors) are outside this network. From previous research showing the evolutionary history of P-loop ion channels[14], these 2 ion channels were diverged from others in an early stage, which makes the distances of evolutionary relatedness too far to be detected. Including the prokaryotic counterparts of these channels would make all these 10 ion channels to be included in a single network. As for the other 6 ligand-gated ion channels, 5 pentameric ligand-gated ion channels can be built into a network. From other studies ([15]) it appears that the connection between the PLGIC's and the P-loop channels is at least very deep in evolutionary history and may have been lost. On the other hand, we find from domain study of P2X a very different story from the PGLIC's. Searching the P2X

domains in InterPro show that they only exist in mammals and are therefore of very recent origin.

Discussions

The homology detection can be achieved by domain-based comparison in addition to sequence-based comparison [15, 16]. Building the evolutionary relationship among functionally-related proteins helps elucidate the structure and function of these proteins. Also, it is believed that protein function should follow largely from domain architecture. Previous researches have shown that domain content could be a reliable basis on the prediction of protein function [10, 17]. This application is essential to the genome projects because functionality can be transferred by computational methods. However, neither the sequenced-based methods nor the domain-based methods work perfectly. One can assist the other one in many kinds of researches.

In this research, the assignment of significance scores on domains by Inverse Domain Frequency (IDF) and Domain Evolutionary Significance (DES) may provide different application. The difference between IDF and DES is the introduction of promiscuity on the weight of the score. Domains with high promiscuity have a strong role in functionality, but are weak indicators of evolutionary relationships because they impart volatility and fast-changing features to a family of sequences [5]. It makes the DES scores more suitable in inferring the evolutionary history than the IDF scores, because we have shown that the cosine similarity scores on DES are more agreeable with OMA orthology definition. In the further development of domain-based research,

the comparisons of proteins in an evolutionary basis should include the consideration of domain promiscuity.

In our case study of 16 human ion channel families, eight out of ten P-loop channels can be connected by domain-based comparison. This is a proof that cosine similarity is a measure that can be used in building the relationship of related proteins. Although two families are outside this network, this has a clear explanation. From the evolution of all P-loop families, these two (inwardly rectified potassium channels and ionotropic glutamate receptor channels) are diverged from the others prior to the divergence between eukaryotes [14]. Since the proteins we used to generate Figure 2 are from humans, the proteins in these two families have diverged so much so that no common domain definition can be found. We propose that including their prokaryotic counterparts in the comparison of domain architectures may reveal a more detailed network and the cosine similarity scores on the network may shed more information about their evolutionary distances.

Up to now, there have been many approaches of domain architecture comparison [6-8]. However, the method of domain comparison has concerns that must be dealt with. First, it relies highly on the accurate and complete domain assignment. If the domain definition is not sufficiently precise or comprehensive, the comparison of domain architecture will be contaminated by false positives (inadequate precision) or false negatives (inadequate comprehensiveness). In one direct comparison it appears that one needs both domain-based and whole sequence-based methods together to achieve the optimum combination of precision and comprehensiveness [15]. Secondly, the interactions of domains need to be further investigated. Here, we only focused on the number of distinct partners (i.e. promiscuity). Some pairs of domains may have co-

occurred at different frequency and the order of the domains may play a minor role in the functionality. Also, the use of cosine similarity scores as measures of evolutionary distance needs further validation. Further improvements of domain assignment and domain architecture comparison will make the domain-based methods even more reliable and more applicable in further research.

An early great discovery using domain-based methods was in Woese and Fox [18]. In that work oligonucleotides (essentially domains) from digestion of ribosomal RNA were compared among organisms, and it was concluded that there are not two but three superkingdoms of life. We believe that domain-based analysis has much to contribute.

Methods

Computations of inverse domain frequency, domain evolutionary significance and cosine similarity scores between domain compositions

First, each domain was assigned a score of significance about its role in determining the functionality and evolution of proteins. Two different measurements were adopted in this research [9]. The first one is Inverse Domain Frequency (IDF). IDF scores were obtained by dividing the number of total domains in a database by the number of the proteins carrying specific domain, and then taking the logarithm of that quotient. The equation is

$$IDF(d) = \log 2 \frac{Nt}{Nd}$$

where Nt is the number of total proteins, Nd is the total number of proteins having a specific domain. This indicates the general importance of a domain because the domains occurred fewer times usually carry more significance. While two proteins are compared with their similarity, domains should be weighted by their significance [7].

The second method is to decrease the impact of highly mobile (promiscuous) domains. These domains happen in a database with extremely high frequency. However, although they might be related to functional relationship, they probably are not involved in orthology. Therefore, Domain Evolutionary Significance (DES) was proposed to penalize the significance if domains are highly mobile. For each domain, the number of distinct partner domains can be counted while retrieving the domain content. The equation is

$$DES(d) = IDF(d) * \frac{1}{Np}$$

where Np is the number of distinct partner domains for domain d. In this measurement, the scores are decreased by the mobility of a domain.

With the domain weight scores, cosine similarity can be calculated between two sets of domains, i.e. the angle between two vectors of the same size. The cosine similarity is defined

$$sim(X,Y) = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}}$$

where X,Y are the domain compositions and x, y are the scores of their member domains. The range cosine similarity scores are from 0, meaning that no shared domain, to 1, meaning the identical domain compositions.

Several perl scripts have been developed for these analyses.

Orthology MAtrix Project (OMA) data retrieval

Orthology MAtrix Project (OMA, http://omabrowser.org) is a database managing orthologs derived from publicly available complete genomes [19]. It produced OMA groups, a subset of orthologous proteins. Within each OMA group, every protein is orthologous to every other protein. Because UniProt database has incorporated OMA information it its databse, OMA groups could be built by parsing the UniProt data files.

Data sets of families of human ion channels

In order to apply the methods we developed, we took the protein sequences from IUPHAR databases (http://www.iuphar-db.org/) [20]. This database manages the protein sequences of human ion channels, which can be divided into ligand-gated ion channels (LGICs) and voltage-gated ion channels (VGLCs). There are 7 types LGICs and 9 types of VGLCs in this databse. We collected all proteins in each ion channel type, parsed their domain compositions and computed the averages of cosine similarity scores for each pairs of these 16 types of ion channels.

Construction of networks by Cytoscape

We constructed the network of cosine similarity among 16 types of ion channels by Cytoscape v.2.6.3. Cytoscape is a software which analyzes networks and provides the

visualization of data (http://www.cytoscape.org/) [21]. The pairs of cosine similarity scores among 16 types of ion channels were input to Cytoscape to generate the network.

References

1. Consortium TU: **The Universal Protein Resource (UniProt) in 2010.** *Nucleic Acids Res* 2010, **38:**D142-148.

2. Schnoes AM, Brown SD, Dodevski I, Babbitt PC: **Annotation error in public databases: misannotation of molecular function in enzyme superfamilies.** *PLoS Comput Biol* 2009, **5:**e1000605.

3. Fong JH, Geer LY, Panchenko AR, Bryant SH: **Modeling the evolution of protein domain architectures using maximum parsimony.** *J Mol Biol* 2007, **366:**307-315.

4. Bjorklund AK, Ekman D, Light S, Frey-Skott J, Elofsson A: **Domain rearrangements in protein evolution.** *J Mol Biol* 2005, **353:**911-923.

5. Basu MK, Poliakov E, Rogozin IB: **Domain mobility in proteins: functional and evolutionary implications.** *Brief Bioinform* 2009, **10:**205-216.

6. Lin K, Zhu L, Zhang DY: **An initial strategy for comparing proteins at the domain architecture level.** *Bioinformatics* 2006, **22:**2081-2086.

7. Song N, Sedgewick RD, Durand D: **Domain architecture comparison for multidomain homology identification.** *J Comput Biol* 2007, **14:**496-516.

8. Song N, Joseph JM, Davis GB, Durand D: **Sequence similarity network reveals common ancestry of multidomain proteins.** *PLoS Comput Biol* 2008, **4:**e1000063.

9. Lee B, Lee D: **Protein comparison at the domain architecture level.** *BMC Bioinformatics* 2009, **10 Suppl 15:**S5.

10. Koestler T, von Haeseler A, Ebersberger I: **FACT: functional annotation transfer between proteins with similar feature architectures.** *BMC Bioinformatics* 2010, **11:**417.

11. Bornberg-Bauer E, Beaussart F, Kummerfeld SK, Teichmann SA, Weiner J, 3rd: **The evolution of domain arrangements in proteins and interaction networks.** *Cell Mol Life Sci* 2005, **62:**435-445.

12. Jin J, Xie X, Chen C, Park JG, Stark C, James DA, Olhovsky M, Linding R, Mao Y, Pawson T: **Eukaryotic protein domains as functional units of cellular evolution.** *Sci Signal* 2009, **2:**ra76.

13. Vogel C, Teichmann SA, Pereira-Leal J: **The relationship between domain duplication and recombination.** *J Mol Biol* 2005, **346:**355-365.

14. Jegla TJ, Zmasek CM, Batalov S, Nayak SK: **Evolution of the human ion channel set.** *Comb Chem High Throughput Screen* 2009, **12:**2-23.

15. Rendon G, Kantorovitz MR, Tilson JL, Jakobsson E: **Identifying bacterial and archaeal homologs of pentameric ligand-gated ion channel (pLGIC) family using domain-based and alignment-based approaches.** *Channels (Austin)* 2011, **5:**325-343.

16. Ger MF, Rendon G, Tilson JL, Jakobsson E: **Domain-based identification and analysis of glutamate receptor ion channels and their relatives in prokaryotes.** *PLoS One* 2010, **5:**e12827.

17. Forslund K, Sonnhammer EL: **Predicting protein function from domain content.** *Bioinformatics* 2008, **24:**1681-1687.

18. Woese CR, Fox GE: **Phylogenetic structure of the prokaryotic domain: the primary kingdoms.** *Proc Natl Acad Sci U S A* 1977, **74:**5088-5090.

19. Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C: **OMA 2011: orthology inference among 1000 complete genomes.** *Nucleic Acids Res* 2011, **39:**D289-294.

20. Sharman JL, Mpamhanga CP, Spedding M, Germain P, Staels B, Dacquet C, Laudet V, Harmar AJ: **IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data.** *Nucleic Acids Res* 2011, **39:**D534-538.

21. Kohl M, Wiese S, Warscheid B: **Cytoscape: software for visualization and analysis of biological networks.** *Methods Mol Biol* 2011, **696:**291-303.

Tables

Table 4-1. - The statistics of cosine similarity scores between pairs of InterPro, Pfam and Gene3D domain compositions

The statistics of cosine similarity scores are listed. The averages are derived from all pairs of domain compositions sharing at least one common domain. Standard deviations are listed in the parenthesis.

Domain definition	# of pairs of cosine similarity scores	Average of similarity scores	Range	
InterPro	195,916,056	0.2046 (0.1476)	Max: 0.99453504171100 Min: 0.01016034232135	On IDF
		0.0370 (0.1175)	Max: 0.99999999884306 Min: 0.00000002367324	On DES
Pfam	19,541,013	0.2766 (0.1614)	Max: 0.99773203916207 Min: 0.00557564398660	On IDF
		0.1159 (0.1.970)	Max: 0.99999999925484 Min: 0.00000000004385	On DES
Gene3D	3,957,161	0.2611 (0.1655)	Max: 0.99766081550983 Min: 0.00565517355953	On IDF
		0.1256 (0.2105)	Max: 0.9999999998778 Min: 0.00000000071581	On DES

Table 4-2. - The average of cosine similarity scores in OMA groups

The statistics of OMA groups and domain composition groups are listed in (A), (B), and (C) respectively for InterPro, Pfam and Gene3D domain fefinitions.

(A)

Of 13,593,921 proteins in UniProt 2011_01:

	with InterPro domians	without InterPro domains	
with OMA defintion	2,481,346	318,913	2,800,259
without OMA definition	8,054,548	2,739,114	10,793,662
	10,535,894	3,058,027	13,593,921

The number of the InterPro domains: 21,091

The number of total domain compositions: 153,165

The number of total OMA groups: 388,214

Of 153,165 domain compositions:

	Without OMA	With one OMA group	With more than one OMA groups
number of domain compositions	112094 (73.19%)	18,984 (12.39%)	22,087 (14.42%)

Of 388,214 OMA groups:

		Without InterPro domain composition		With one InterPro domain composition		With one InterPro domain composition *		With more than one InterPro domain composition		With more than one InterPro domain composition *	
number of OMA groups		74,836 (19.28%)		210,816 (54.30%)		13,040 (3.36%)		86,585 (22.30%)		2,937 (0.76%)	
Average of cosine similarity scores on IDF	Average of numbers of domain compositions	NA	0	1	1	1	1	0.7348 (0.1684)	2,43 (1.01)	0.4338 (0.3178)	2.57 (1.24)
Average of cosine similarity scores on DES		NA		1		1		0.5771 (0.3389)		0.3629 (0.3650)	

*: Some proteins within this group are without InterPro domain compositions. When computing the cosine similarity scores, they were excluded.

Table 4-2. (Continued)

(B)
Of 13,593,921 proteins in UniProt 2011_01:

	with Pfam domians	without Pfam domains	
with OMA defintion	2,371,337	428,922	2,800,259
without OMA definition	7,606,241	3,187,421	10,793,662
	9,977,578	3,616,343	13,593,921

The number of the Pfam domains: 11,464
The number of total domain compositions: 77,561
The number of total OMA groups: 388,214
Of 77,561 domain compositions:

	Without OMA	With one OMA group	With more than one OMA groups
number of domain compositions	54,192 (69.87%)	9,725 (12.54%)	22,087 (18.11%)

Of 388,214 OMA groups:

		Without Pfam domain composition		With one Pfam domain composition		With one Pfam domain composition *		With more than one Pfam domain composition		With more than one Pfam domain composition *	
number of OMA groups		92,274 (23.76%)		234,975 (60.53%)		18,806 (4.84%)		39,121 (10.08%)		3,038 (0.78%)	
Average of cosine similarity scores on IDF	Average of numbers of domain compositions	NA	0	1	1	1	1	0.6578 (0.2226)	2,25 (0.67)	0.3547 (0.3319)	2.46 (0.99)
Average of cosine similarity scores on DES		NA		1		1		0.5933 (0.3246)		0.3275 (0.3453)	

*: Some proteins within this group are without Pfam domain compositions. When computing the cosine similarity scores, they were excluded.

Table 4-2. (Continued)

(C)
Of 13,593,921 proteins in UniProt 2011_01:

	with Gene3D domians	without Gene3D domains	
with OMA defintion	1,002,563	1,797,696	2,800,259
without OMA definition	3,477,419	7,316,243	10,793,662
	4,479,982	9,113,939	13,593,921

The number of the Gene3D domains: 1147
The number of total domain compositions: 14,451
The number of total OMA groups: 388,214
Of 14,451 domain compositions:

	Without OMA	With one OMA group	With more than one OMA groups
number of domain compositions	10,104 (69.92%)	1,817 (12.57%)	2,530 (17.51%)

Of 388,214 OMA groups:

		Without Gene3D domain composition		With one Gene3D domain composition		With one Gene3D domain composition *		With more than one Gene3D domain composition		With more than one Gene3D domain composition *	
number of OMA groups		259,867 (66.94%)		100,043 (25.77%)		12,096 (4.84%)		14,879 (3.83%)		1,329 (0.34%)	
Average of cosine similarity scores on IDF	Average of numbers of domain compositions	NA	0	1	1	1	1	0.7105 (0.1609)	2,20 (0.58)	0.4768 (0.3195)	2.46 (0.99)
Average of cosine similarity scores on DES		NA		1		1		0.6811 (0.2420)		0.4618 (0.3302)	

*: Some proteins within this group are without Pfam domain compositions. When computing the cosine similarity scores, they were excluded.

Figures

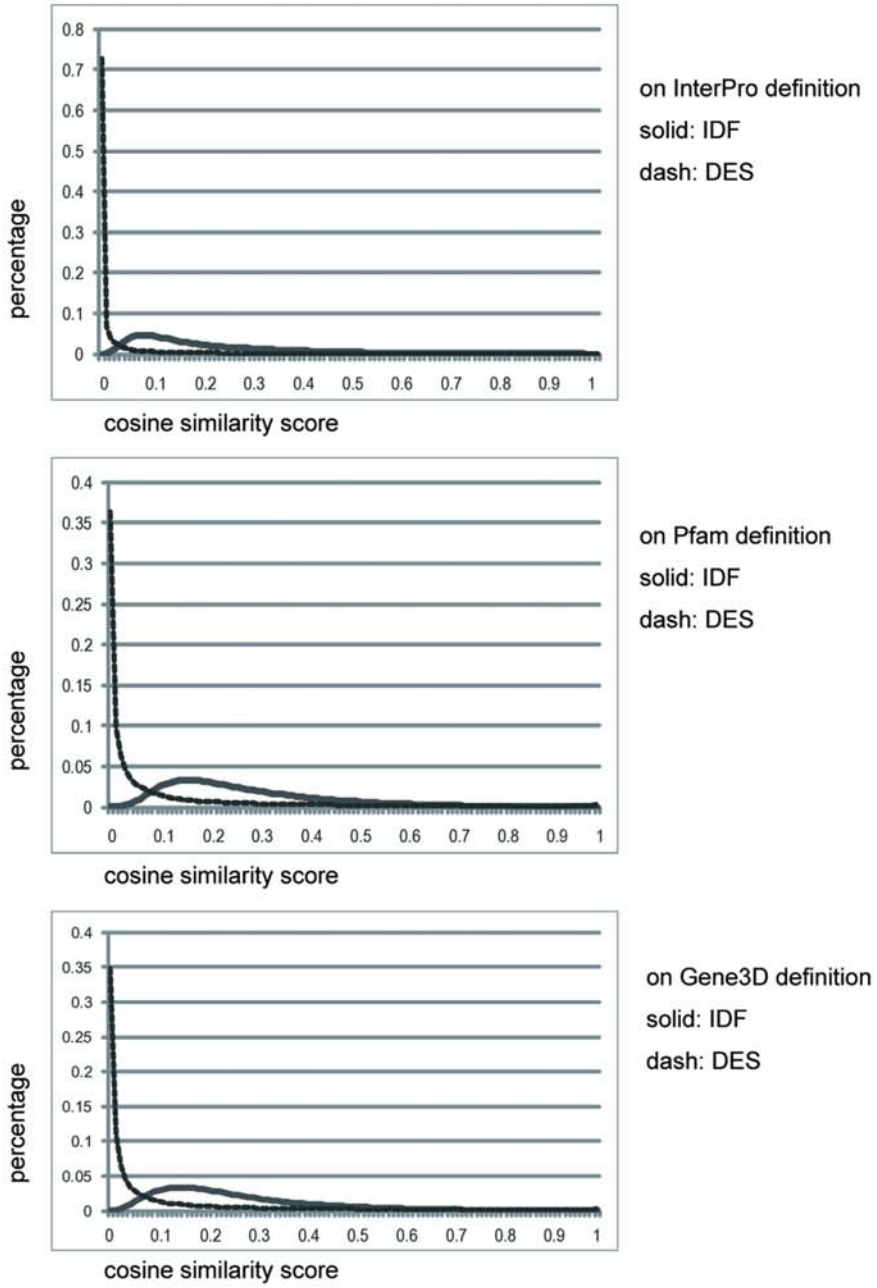


Figure 4-1. - the distributions of cosine similarity scores for InterPro, Pfam and Gene3D domain definitions
The top is the distribution of non-zero cosine similarity scores for InterPro domain definition. The middle is the distribution of non-zero cosine similarity scores for Pfam domain definition. The bottom is distribution of non-zero cosine similarity scores for Gene3D domain definition.

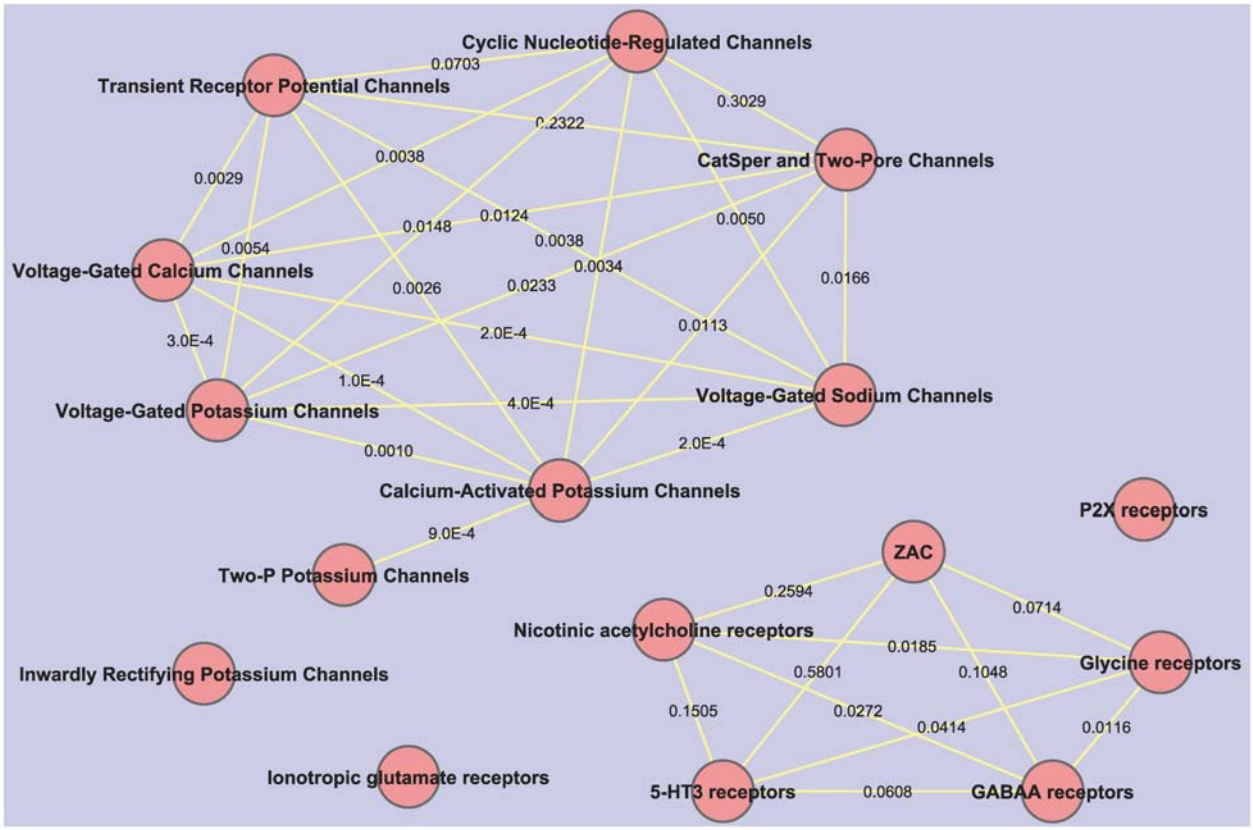


Figure 4-2. - the network of cosine similarity scores on 16 types of ion channels
This figure is the network of 16 families of human ion channels. Connection between nodes denotes the sharing of domains between their proteins. The values on edges are the averages of cosine similarity scores (based on DES) between two ion channel families.

CHAPTER 5:

FUTURE WORK

The importance of building a model of domain evolution

Protein evolution happens at different scales of event. On the smallest one, bases and the corresponding amino acids are the basic units of changes; on the large one, domains are the leading roles. There have been many researches devoted to interpret the domain content in proteomes [1]. The extant repertoire of domain architectures, which are frequently regarded as a fundamental level of protein function complexity, were derived from fusion, fission, recombination, splicing, as well as slow creations of new domains from extant domains. Increasing domain combinations by fusion is the major force to expand the repertoire [2]. Extant domain combinations are the result of selective forces, making them to remain in the repertoire. Some promiscuous domains may play an important role in the process of domain rearrangement. Moreover, an organism's complexity may be more related to the number of distinct multi-domain architectures than to the genome size [3]. Although the rearrangement process is not clear yet, there is urgent need to build a model which can illustrate the evolutionary pathways by which extant domain architectures may have evolved.

Preliminary thoughts about the model

This model should include the following features, but not limited to:

- Discrete-time stochastic model: A stochastic process is a probability model that describes the evolution of a system. Discrete-time defines the stages along the history of domain evolution. The probabilities used in this model could be inferred from the extant proteomes.
- Hidden Markov model: A hidden Markov model describe a series of states which can represent the repertoires at different time of evolution. The transitions between these states mimic the processes of domain rearrangements.
- Domain definition: From our experience, InterPro is the most comprehensive domain database. Although this model can be evaluated under different domain definition, it will start with InterPro domain definition.
- Order of domains, neighbor effect, and position effect: In our research thus far, the order of domains,the neighbor effect, and the position effect are not considered. To gain accuracy, this model should take the order of domains and the neighbor effect (domains occurred together at high frequency) into its simulation.
- Maximum parsimony principle: The inference from observed data that requires the least evolutionary change.

Although we cannot obtain true ancestral protein architectures, this model will be used in the applications described in the next section.

The application of the domain evolution model

Once the model is validated, several applications could be made:

- The evolutionary history could be tracked in terms of domain rearrangement. This process represents the adaption and selection of various organisms. The illustration of the process provides the foundation of deciphering the history of life.
- The functional complexity of species could be accounted by the parameters of the model. The speed of evolution on a specific species might be approximately predicted, especially for the fast-changing prokaryotes.
- The research of proteomics could proceed in a new direction, which is focused on the modularity of proteins. This may facilitate the discovery of new biomarker and protein drugs.

References

1. Buljan M, Bateman A: The evolution of protein domain families. *Biochem Soc Trans* 2009, 37:751-755.

2. Fong JH, Geer LY, Panchenko AR, Bryant SH: Modeling the evolution of protein domain architectures using maximum parsimony. *J Mol Biol* 2007, 366:307-315.

3. Babushok DV, Ostertag EM, Kazazian HH, Jr.: Current topics in genome evolution: molecular mechanisms of new gene formation. *Cell Mol Life Sci* 2007, 64:542-554.