Evolution of a protein domain interaction network*

Gao Li-Feng(高丽锋)[a], Shi Jian-Jun(石建军)[b], and Guan Shan(官 山)[b]†

[a] *Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing* 100081, *China*

[b] *College of Physics Science and Technology, Yangzhou University, Yangzhou* 225002, *China*

(Received 29 March 2009; revised manuscript received 28 April 2009)

In this paper, we attempt to understand complex network evolution from the underlying evolutionary relationship between biological organisms. Firstly, we construct a Pfam domain interaction network for each of the 470 completely sequenced organisms, and therefore each organism is correlated with a specific Pfam domain interaction network; secondly, we infer the evolutionary relationship of these organisms with the nearest neighbour joining method; thirdly, we use the evolutionary relationship between organisms constructed in the second step as the evolutionary course of the Pfam domain interaction network constructed in the first step. This analysis of the evolutionary course shows: (i) there is a conserved sub-network structure in network evolution; in this sub-network, nodes with lower degree prefer to maintain their connectivity invariant, and hubs tend to maintain their role as a hub is attached preferentially to new added nodes; (ii) few nodes are conserved as hubs; most of the other nodes are conserved as one with very low degree; (iii) in the course of network evolution, new nodes are added to the network either individually in most cases or as clusters with relative high clustering coefficients in a very few cases.

Keywords: complex network, protein domain, network evolution
PACC: 0590, 8710

1. Introduction

The modelling of complex systems as networks has revealed a set of generic organizing principles;[1−3] most of these systems are small world, scale-free and modular, resulting in a hierarchical architecture. It is believed that the emergence of the ubiquitous topological properties of real networks originated from network evolution, and enormous evolutionary models were put forward to explain these phenomena,[3−6] including the Barabasi–Albert (BA) model[3] whose two key points are growth and preferential attachment, and the adaptive network model[4] in which both topology and dynamics of the network co-evolved with each other. For further understanding of network evolution, new ideas are expected from more empirical systems.

A variety of biological systems, including protein domain interaction network and metabolic network, have been found to exhibit similar large-scale traits as those found in other disciplines.[7−11] At the same time, we have always believed that there exists an underlying evolutionary relationship between organisms, since Charles Darwin's 'The Origin of Species' was published. Therefore, biological systems may be good candidates for research on complex network evolution.

Proteins are composed of evolutionarily conserved units called domains. Novel yet specific combinations of domains are essential for creating diversity in proteins. The sequential order of domains in a protein sequence is known as its protein domain architecture. Advances in genome sequencing have generated hundreds of completely sequenced genomes, and the map between protein domain architectures and known protein sequences are covered by proteome annotation.[12−15] Since protein domains are evolutionarily conserved units of protein, interactions between protein domains are widely studied and relative achievements are accumulated into databases.[16,17] This situation makes it possible to construct a protein domain interaction network for completely sequenced organisms, and to analyse the evolutionary process of the network based on the organism evolutionary relationship.

In this paper, we use the Pfam domain[14] interaction network as our research object and attempted to understand complex network evolution based on the underlying evolutionary relationship between biological organisms. In Section 2, we construct a Pfam domain interaction network for each of the 470 completely sequenced organisms depending on both the annotation of co-occurrence of Pfam domains within a protein[12] and databases[16,17] with information on the Pfam domain interaction relationship. This helps us to correlate each organism with a specific Pfam domain interaction network; in Section 3, we infer

*Project supported by the National Natural Science Foundation of China (Grant Nos. 70671089 and 30871521), and the State Key Program of National Natural Science of China (Grant No. 10635040).

†Corresponding author. E-mail: dhq2gs@yahoo.com.cn

© 2010 Chinese Physical Society and IOP Publishing Ltd

the evolutionary relationship of these organisms with the nearest neighbour joining method in which organisms are denoted by a vector with its component being assigned to 1 or 0 depending on the existence or nonexistence of the corresponding Pfam domain, and the Hamming distance[18] between vectors was used to construct the evolutionary tree of organisms. Our results show that these 470 organisms are correctly divided into Archaea, Bacteria, and Eukaryota, and most of the major groups within these kingdoms were sorted satisfactorily. Based on the results of Sections 2 and 3, in Section 4, we obtain 3 sub-networks for Archaea, Bacteria and Eukaryota respectively. They consist of nodes shared by all organisms of each kingdom and positioned at the roots of the evolutionary trees; in Section 5, we analyse the evolutionary process of the Pfam domain interaction network and attempt to understand how and where new nodes are added to the network; finally, in Section 6, we draw a conclusion from our results.

2. The Pfam domain interaction network of an organism

In order to construct a Pfam domain interaction network for each organism, firstly we have to set up the mutual interaction relationship of Pfam domains. Actually, we used two kinds of data resources for this purpose. The first one is a database about tje Pfam domain interaction relationship obtained by experiment or prediction, including DIMA[16] and DOMINE.[17] These two databases offered us 28874 interactions between 4958 Pfam domains and 20513 interactions between 4036 Pfam domains, respectively; the second one is the co-occurrence of Pfam domains[19] within a protein which is parsed from domain architectures of proteins annotated in the uniprot database.[12] Domains co-occurring within a protein are considered to be a fully connected graph in which the edge between two nodes represents a mutual interaction relationship between two domains. This data resource offered us 25386 interactions between 5674 Pfam domains. By merging these data resources together and removing self-interactions from the final dataset, we obtained 61373 interactions between 6968 Pfam domains (Table 1). Obviously, in our dataset, the interaction relationship not only means physical interactions observed in experiments, but also computationally predicted interactions. However, compared with the number of possible mutual interaction relationships(6968×6968), we find that our data still comprise a very small subset.

Table 1. Data resources for Pfam domain interaction.

	DIMA	Domine	uniprot	total
domain No.	4958	4036	5674	6968
interaction No.	28874	20513	25386	61373

On the other hand, several hundred organisms distributed between viruses, Archaea, Bacteria and Eukaryota are fully sequential and their protein annotation is maintained in the uniprot database.[12] These datasets were further filtered according to the following steps: (i) all of the viruses were discarded since their genomes are very small and they contain very few Pfam domains; (ii) only a representative example of organisms that belong to the same genus is reserved since their Pfam domain sets are similar; (iii) organisms whose Pfam domain numbers are much smaller than the others are removed from our dataset, since their evolutionary distances to others are too far and are arranged in an isolated branch of the evolutionary tree. Particularly, for Archaea, except for 1 organism whose Pfam domain number is 347, all of the others have more than 730 domains; for Eukaryota, except for 4 organisms whose Pfam domain numbers are less than 323, all the others have more than 750 domains; for Bacteria, their domain numbers vary from a few to thousands. Taking account of these observations, we chose number 700 as our threshold and organisms which contain less than 700 Pfam domains were discarded. Finally, we collected 470 organisms, and 46 of them belong to Archaea, 368 to Bacteria, and 56 to Eukaryota.

Based on the mutual interaction relationship of the Pfam domain, we constructed a Pfam domain interaction network for each of the 470 organisms. Our research shows that at least 95% of the domains of each organism are involved in a giant connected graph. Then, after eliminating domains isolated from the giant graph, we correlated each organism with a specific network whose nodes and edges represent Pfam domains and interactions between domains respectively. For an overall survey, we calculated node degree distribution for each organism. By dividing all organisms into three kingdoms and normalizing node degree to [0, 1], all of the node degree distributions of the same kingdom are arranged in the same plane (Fig. 1). At the same time, as a reference system, an artificial network was constructed for each kingdom by merging all organisms' networks together, and both of its node degree distribution and linear fitted line overlap on the figure. Generally, for each organism, their node degree distribution fluctuated around a power law distribution.

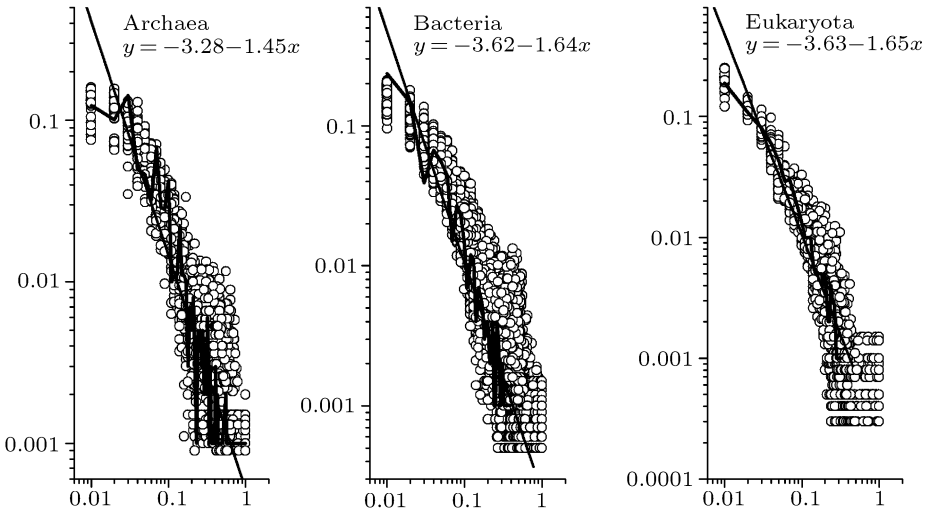


Fig. 1. Organisms' degree distributions are shown as scattered circles instead of curves for clarity. The zigzag line is the node degree distribution of the artificial network, and both its linear fitted line and equation are also shown in the figure. The x coordinate is the normalized node degree. The y coordinate is degree distribution.

3. Evolutionary relationship between organisms

In the literature, the evolutionary relationship between organisms is illustrated as a tree called the phylogenetic tree, and its construction relies on using mathematical methods to infer the past from features of contemporary organisms. Traditionally, evolutionary tree construction was based on the analysis of morphological characters; the main groups of animals and higher plants have been classified fairly easily by using these methods. However, the power of this approach is restricted in micro-organisms since very limited morphological features are available. The introduction of the use of molecular data in phylogenetics initiated molecular phylogeny of prokaryotes.[20] Individual genes with a considerable degree of conservation across all organisms were extensively used for evolutionary tree construction of micro-organisms. However, as more genes were analysed, contradiction instead of refinement between phylogenies based on comparisons of single genes was revealed.[21] With the advent of the era of complete genome sequences a variety of approaches based on whole genome features for determining the evolutionary history of organisms were put forward (see Ref. [22] for review).

In our work, since each organism is correlated with a specific Pfam domain interaction network, it is naturally to represent each of them by the set of Pfam

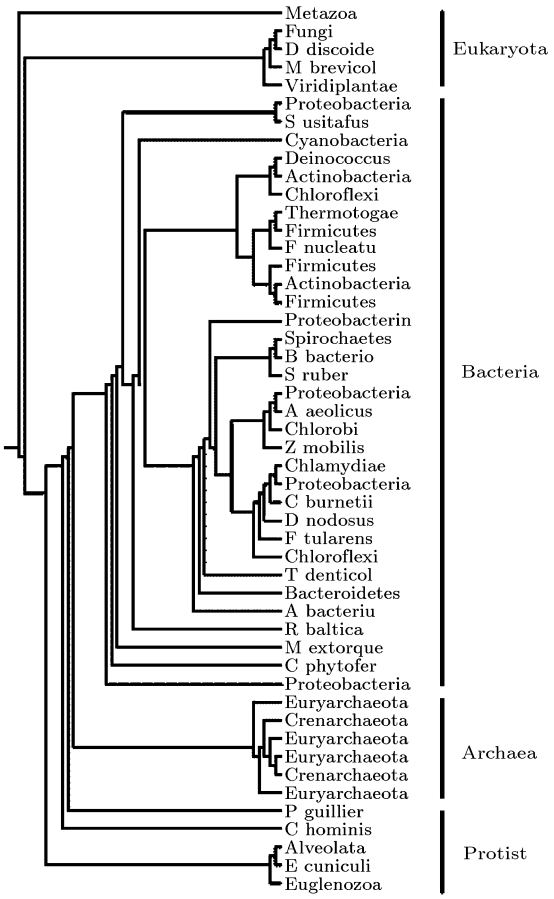


Fig. 2. A phylogenetic tree where its leaves have been merged into phyla as far as possible. The first letter of the genus and the first 8 letters of the species are composed together to represent an organism.

domains contained, and to measure their evolutionary relationship by differences between Pfam domain sets. Intuitively, the closer two organisms are, the more Pfam domains they share. Therefore, we constructed a vector with N components: $V = [D_1, D_2, \ldots, D_N]$ where N equals the total number of Pfam domains contained in all organisms, and D_i represents the ith Pfam domain. Then, each organism is assigned with a specific vector with its component which is equal to 1 if the corresponding Pfam domain is contained in the organism, or 0 if not. Different kinds of distance between organisms can be calculated on these vectors. For simplicity, we use the Hamming distance definition[18] to calculate the evolutionary distance between two organisms, which is defined as: $d(\boldsymbol{V}_x, \boldsymbol{V}_y) = \sum_{i=1}^{N} |V_{xi} - V_{yi}|$, where $\boldsymbol{V}_x$ and $\boldsymbol{V}_y$ are vectors assigned to organism x and y respectively, their vector components are denoted as $\boldsymbol{V}_{xi}$ and $\boldsymbol{V}_{yi}$, and the summation goes through all Pfam domains of a kingdom. Then, we obtained a distance matrix between organisms, and the standard neighbour-joining program of the PHYLIP package (http://www.phylip.com/) can be used to generate the evolutionary tree.

To show the evolutionary tree of all organisms in one figure, we tried to merge the leaves (which are organisms) of the tree into phyla as far as possible and obtained Fig. 2. We were encouraged by two observations: (i) at the granularity of phylum, except for a few individuals that are the only members of a phylum, all organisms were correctly sorted into different phyla; (ii) at the granularity of kingdom, all organisms were correctly classified into Archaea, Bacteria, and Eukaryota (marks at the right most side of the figure; protists, which are traditionally classified into Eukaryota, are listed separately for clarity). Something that needed to be analysed further is: some phyla were separated into different branches of the tree, especially phyla in Bacteria and Archaea. For example, proteobacteria is separated into 5 branches.

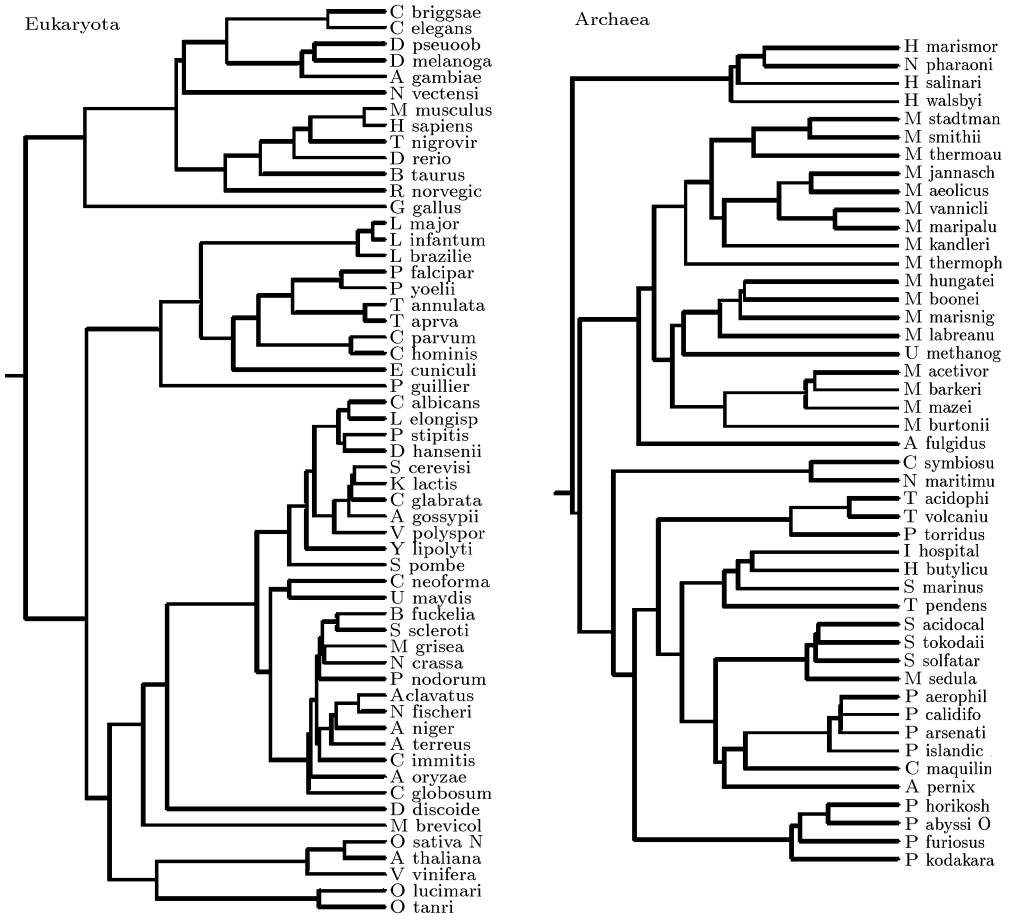


Fig. 3. Evolutionary tree of Archaea (46 organisms) and Eukaryota (56 organisms). Organisms' names are abbreviated in the same way as those in Fig. 2.

At the organism level, the phylogenetic tree for 46 organisms of Archaea and 56 organisms of Eukaryota is shown in Fig. 3 (for 368 organisms of Bacteria, the evolutionary tree is too large to be listed here). Although not every organism appears exactly at its classical evolutionary position in these trees, most of the major groups are satisfactorily sorted within these kingdoms (note: the first letter of the genus and the first 8 letters of the species are composed together to represent an organism. Organisms coming from the same branch and with the same first letter almost certainly belong to the same genus. This observation was checked by us one by one). The evolutionary relationship between organisms is used for our further analysis of network evolution.

4. Conserved sub-network structure

In our scheme of evolutionary tree construction, we assumed that the Pfam domain interaction networks of different organisms evolve from a common ancestor who is represented by the root of the evolutionary tree, and leaves of the tree correspond to organisms, and is characterized by a specific Pfam domain interaction network; the branch points of the tree are actually sub-networks consisting of Pfam domains shared by its two branches. Therefore, the root of the tree consists of Pfam domains shared by all organisms and is conserved in evolution. For Archaea, 235 Pfam domains are shared by 46 organisms; for bacteria, 33 domains are shared by 368 organisms; for Eukaryota, 142 domains are shared by 56 organisms; and for the three kingdoms, 14 domains are shared by all 470 organisms. To detect the structure of the conserved sub-network, we calculated their clustering coefficients[1] and analysed their k-clique community structures using CFinder.[23] The results listed in Table 2 show that the conserved domains make up a sub-network with a relatively high clustering coefficient, and the majority of them are partners of k-clique communities. For Pfam domains conserved by all organisms, it is reasonable to expect that they play key roles in biological functions. As an example, the functional annotations about the 14 Pfam domains conserved in 470 organisms are listed in Table 3 (adapted from (http://pfam.sanger.ac.uk/). Clearly, they are components of ATPase, acetytransferase, helicase, hydrolase, elongation factor, aminotransferase, etc., which act as catalysts in biological reactions, such as energy transformation, protein synthesis and DNA transcription.

Table 2. Features of the conserved Pfam domain sub-network.

	shared domain	CC	domain in COMM
Archaea	235	0.23	207
Bacteria	33	0.35	30
Eukaryota	142	0.21	125
ABE	14	0.16	11

CC is the abbreviation of clustering coefficient.

COMM is the abbreviation of k-clique community.

ABS is the abbreviation of Archaea, Bacteria and Eukaryota.

Table 3. Description of the 14 Pfam domains conserved in 470 organisms.

PF00004	ATPase family associated with various cellular activities
PF00583	acetyltransferase (GNAT) family
PF00009	Elongation factor Tu GTP binding domain
PF00118	TCP-1/cpn60 chaperonin family
PF00271	helicase conserved C-terminal domain
PF00702	haloacid dehalogenase-like hydrolase
PF00266	aminotransferase class-V
PF00270	DEAD/DEAH box helicase
PF02518	histidine kinase-, DNA gyrase B-, and HSP90-like ATPase
PF03144	elongation factor Tu domain 2
PF00679	elongation factor G C-terminus
PF00557	metallopeptidase family M24
PF07690	major facilitator superfamily
PF00005	ABC transporter

5. Network evolution

In order to understand how new nodes are added to the network, at each branch point of the evolutionary tree, we calculated the clustering coefficient[1] of the sub-network composed of new added nodes. For example, at branch point i of the evolutionary tree, network $N_i(V_i,\ E_i)$ evolved to $N_{i+1}(V_{i+1},\ E_{i+1})$ and $N_{i+2}(V_{i+2},\ E_{i+2})$, where V_x and E_x represent sets of network node and edge respectively, and the new added nodes are denoted by $(V_{i+2}-V_i)$ and $(V_{i+1}-V_i)$. To measure how new nodes are added to networks, we calculated the clustering coefficients of the two sub-networks formed by new added nodes. This procedure is repeated at each branch point of the evolutionary tree. Finally, we obtained a distribution about the clustering coefficient of new added sub-networks, and the results are shown in Fig. 4. For Archaea, Bacteria and Eukaryota respectively, network evolution occurred at 90, 734 and 110 branch points of the phylogenetic tree. However, in only 11 branch points of Archaea, 82 of Bacteria, and 5 of Eukaryota, the clustering coefficients of the new added sub-network are greater than 0.1 (the corresponding probabilities are $P_{\rm Archaea}$ (CC>0.1)=0.12, $P_{\rm Bacteria}$ (CC>0.1)=0.11 and $P_{\rm Eukaryota}$ (CC>0.1)=0.04, respectively). In these cases, new nodes are able to be added to the network as clusters. However, in most cases, the clustering coefficient of new added sub-networks is approximately 0, and the majority of the nodes are added to the network individually.

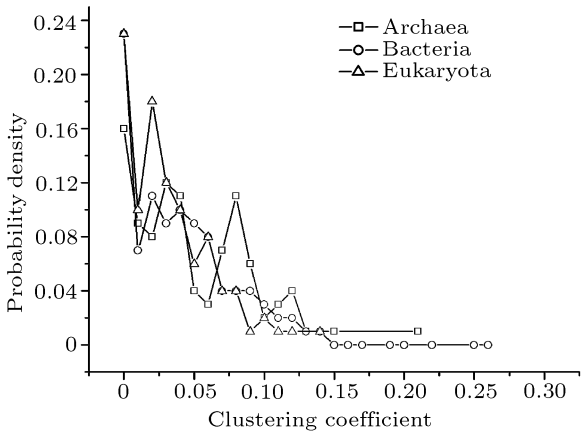


Fig. 4. Probability distribution of the clustering coefficient of new added sub-networks in network evolution.

Since the addition of new nodes to the network will lead to variation of node degree, we use degree variation to locate nodes, to which new nodes were added. Our analysis started from the root of the evolutionary tree. Initially, the sub-network at the root is parsed out and denoted as N_0 (V_0, E_0), where V_0 and E_0 represent sets of node and edge of the network respectively. Then each node is assigned to a value $k(v_i,\ 0)$, which is the degree of node v_i at the branch point (root of the evolutionary tree), $i = 1, \ldots n$. In descending order, these nodes are arranged into a vector $[k(v_0,\ 0)k(v_1,\ 0), \ldots,\ k(v_n,\ 0)]$ with $k(v_0,\ 0) \geq k(v_1, 0) \geq \ldots \geq k(v_n, 0)$. Because of sharing by all organisms (each organism corresponds to a leaf of the tree), nodes in V_0 will appear at each branch point of the tree, and their node degrees vary when new nodes are added. Suppose that, at branch point t of the phylogenetic tree, degrees of nodes in V_0 are $[k(v_0,\ t),\ k(v_1t), \ldots, k(v_nt)]$, and for one of its two branches, their degrees are $[k(v_0,\ t+1),\ k(v_1t+1), \ldots, k(v_nt{+}1)]$, then $[k(v_0,\ t){-}k(v_0,\ t{+}1), k(v_1, t){-}k(v_1, t+1),\ \ldots,\ k(v_n, t) - k(v_n, t+1)]$ were recorded to show where new nodes were added to. For example, if the absolute value of $k(v_1, t) - k(v_1, t+1)$ is the biggest of all variations, it means that most new nodes were added to node v_1. This process was repeated at each branch point, and the average degree variation was finally used to indicate the place where new nodes were most likely be added. Tracing the network evolution step by step, recording all degree variations together, and calculating the average degree variation for each conserved node, we obtained Fig. 5. Obviously, the bigger the node degree is, the bigger its degree varies, and the more new nodes are added to. This phenomenon is consistent with the preferential attachment principle of the BA model.[2] However, for nodes with lower degree, their average node degree variations are rather small, which implies that they prefer to maintain their connectivity in evolution. This observation is different from the preferential attachment principle which claims that node degree should be proportional to its life span. From the aspect of network evolution, we concluded that hubs tend to preserve their role as a hub by connecting with more new added nodes, and nodes with lower degree survived in evolution by keeping their connectivity stable.

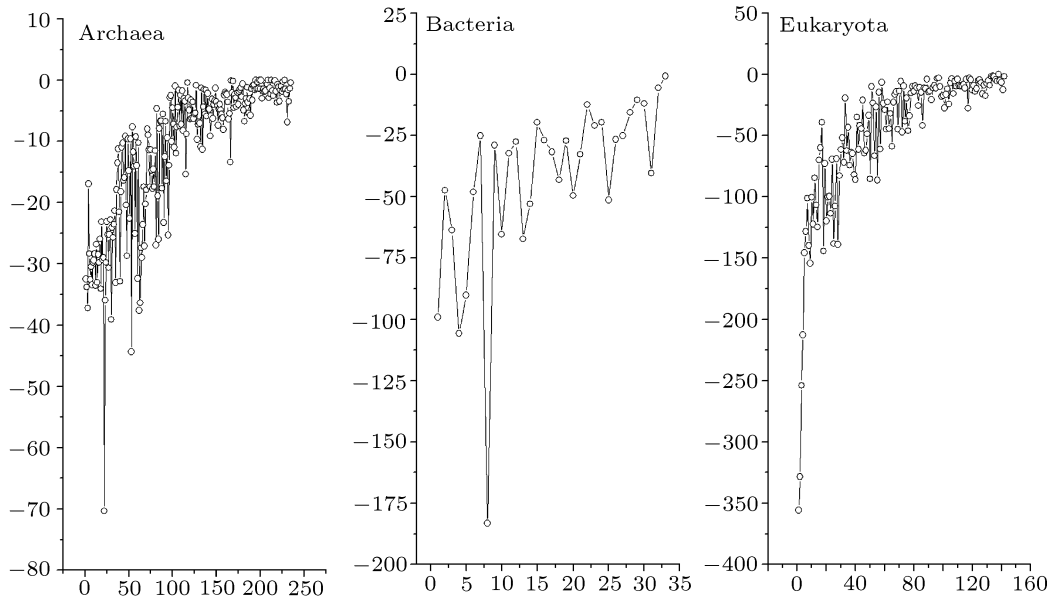


Fig. 5. Average degree variation of conserved nodes in a sub-network. The x axis represents nodes of the conserved sub-network sorted by degrees in descending order. The y axis represents average degree variation. See text for details.

The network structure of each organism can be regarded as a special evolutionary result which is achieved by the sub-network N_0 (V_0, E_0) in evolution, and all the network structures constitute a sample space, in which each sample gives us a distribution showing how nodes in V_0 are distributed at certain degree values. In particular, there are 46 samples for Archaea, 368 for Bacteria and 46 for Eukaryota. By normalizing the node degree of each organism to a value between 0 and 1, and dividing the interval [0, 1] into 100 bins, we obtained an average distribution about the frequency of the conserved node at a certain node degree level (Fig. 6). The high probabilities at lower degree levels imply that most conserved nodes tend to maintain their low degree in network evolution; few nodes keeping their role as hubs in network evolution imply that when new nodes are added they are added to hubs. This observation is consistent with the result obtained above.

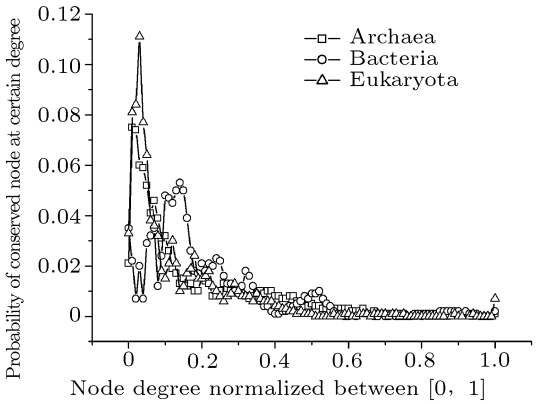


Fig. 6. Distribution of conserved nodes at a certain degree level.

6. Conclusion

The evolutionary process of complex networks can help us in many aspects of understanding the mechanisms of complex systems. However, valuable systems have evolved over a long history and their evolutionary process is impossible to trace. Therefore we have to infer their evolutionary history depending on contemporary observation. The underlying evolutionary relationship between biological organisms makes biological systems good candidates for research on complex system evolution. In this paper, we attempted to analyse the evolution of a Pfam domain interaction network based on an inferred evolutionary tree of organisms, and obtained some interesting observations, including the conserved sub-network structure, the fact that nodes with lower degree in a conserved sub-network prefer to maintain their connectivity invariant in evolution, that hubs of the conserved sub-network are preferentially attached when new nodes are added to the network, and that, in most cases, new nodes are added to a network individually, and in very few cases as clusters.

Initially, we expected that the deviation of network structures between organisms would help us un-

derstand their dissimilar functions. Therefore, analysis of the variation of network structures in evolution is available to uncover functional modules of the network. However, our observations on how new nodes are added to the network in evolution give us a negative response. Taking into account that our dataset of the Pfam domain interaction relationship is only a very small subset of the potential interactions, we find that our Pfam domain interaction networks for each organism are far from complete. Alternatively, maybe the recognition of biological function modules can only be dealt with on a larger scale instead of an isolated Pfam domain interaction network.

Although restricted by the current stage of biological research, we still believe that it is valuable to investigate biological network structures in the history of biological evolution.

References

[1] Watts D J and Strogatz S 1998 *Nature* **393** 440

[2] Albert R and Barabási A L 2002 *Rev. Mod. Phys.* **74** 47

[3] Barabási A and Albert R 1999 *Science* **286** 509

[4] Gross T and Blasius B 2008 *J. R. Soc. Interface* **5** 259

[5] Yang H X, Gao K, Han X P and Wang B H 2008 *Chin. Phys.* B **17** 2759

[6] Guo R Y and Huang H J 2008 *Chin. Phys.* B **17** 1698

[7] Jeong H, Tombor B, Albert R, Oltvai Z and Barabási A L 2000 *Nature* **407** 651

[8] Fell D and Wagner A 2000 *Nat. Biotechnol* **189** 1121

[9] Wuchty S 2001 *Mol. Biol. Evol.* **18** 1694

[10] Apic G, Gough J and Teichmann S 2001 *J. Mol. Biol.* **310** 311

[11] Wuchty S 2002 *Proteomics* **2** 1715

[12] Wu C H, Apweiler R, Bairoch A, Natale D A, Barker W C, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin M J, Mazumder R, O'Donovan C, Redaschi N and Suzek B 2006 *Nucleic Acids Res.* **34** D187

[13] Letunic I, Copley R R, Pils B, Pinkert S, Schultz J and Bork P 2006 *Nucleic Acids Res.* **34** D257

[14] Finn R D, Mistry J, Schuster-Bockler B, Griffiths-Jones S Hollich V, Lassmann T, Moxon S, Marshall M and Khanna A 2006 *Nucleic Acids Res.* **34** D247

[15] Mulder N J, Apweiler R, Attwood T K, Bairoch A, Bateman A Binns D, Bork P, Buillard V and Cerutti L 2007 *Nucleic Acids Res.* **35** D224

[16] Pagel P, Oesterheld M, Stümpflen V and Frishman D 2006 *Bioinformatics* **22** 997

[17] Raghavachari B, Tasneem A, Przytycka T and Jothi R 2008 *Nucleic Acids Res.* **36** D656

[18] Hamming R W 1950 *Bell Syst. Tech. J.* **26** 147

[19] Wuchty S and Almaas E 2005 *BMC Evol. Biol.* **23** 24

[20] Woese C R and Fox G E 1977 *Proc. Natl. Acad. Sci. USA* **74** 5088

[21] Delsuc F, Brinkmann H and Philippe H 2005 *Nat. Rev. Genet.* **6** 361

[22] Brocchieri L 2001 *Theor. Popul. Biol.* **59** 27

[23] Palla G, Derényi I, Farkas I and Vicsek T 2005 *Nature* **435** 814