WUJNS
Wuhan University Journal of Natural Sciences

Vol.11 No.3 2006 577-580

Article ID: 1007-1202(2006)03-0577-04

Duplication and Combination of P-Loop Containing Nucleotide Triphosphate Hydrolases Superfamily

□ SANG Jianping[1,2], GUAN Wei[1], ZOU Xianwu[1]

1. School of Physics Science and Technology, Wuhan University, Wuhan 430072, Hubei, China;
2. Department of Physics, Jianghan University, Wuhan 430056, Hubei, China

Abstract: In a genome the set of proteins are formed by duplication and combination of domain superfamilies. P-loop containing nucleotide triphosphate (NTP) hydrolases superfamily is massively duplicated and has the most different partner superfamilies among archaea, bacteria and eukarya. Here, we study the distributions of duplication and combination of p-loop containing NTP hydrolases superfamily in 169 completed genomes. When the total number of domains in a genome is larger, duplication and combination partners of p-loop containing NTP hydrolases are more. This phenomenon is more obvious in metazoa. The distributions of abundance and combination of partners relate to the functions of the protein. Those distributions in metazoa are very different from those in other kingdoms because of complexity of metazoa. Finally the relationship between duplication and combination of p-loop containing NTP hydrolases superfamily in different genomes is described. It fits a power law.

Key words: p-loop containing NTP hydrolases; combination; duplication; genome

CLC number: Q 617; Q 811.4

Received date: 2005-11-19

Foundation item: Supported by the National Natural Science Foundation of China (10374072)

Biography: SANG Jianping (1959-), male, Professor, Ph.D., research direction: soft condensed matter. E-mail: jpsang@whu.edu.cn

0 Introduction

Protein domains represent the basic evolutionary units that form protein. Examination of their properties provides a key to understanding the evolution of proteomes and of the organism′s complexity[1]. P-loop containing nucleotide triphosphate (NTP) hydrolyses superfamily is an important part of protein repertoire, whose members can function as kinases with very different specificities, as different kinds of motor proteins, and as batteries to drive reactions through conformational change[2].

Duplication is very important at the level of domains. At least 58% of the domains in Mycoplasma and 98% of domains in human[3] are duplicated[4]. The domains of different superfamilies are duplicated to varying extent and the distribution of superfamily sizes in genomes follows a power law[5]. It means that there exist a few highly abundant superfamilies[4]. It seems that the distribution of superfamily size is mainly the result of selection for useful functions rather than the stochastic process[2].

Proteins formed by combinations of domains are particularly abundant in eukaryotes. They occupy more than 80% of all matched protein sequences in eukaryotes. In prokaryotes these kinds of proteins are somewhat less abundant, but still occupy the majority of matched protein sequences (about 65%)[4]. For a few superfamilies, the member of a superfamily is the combination of many domains belonging to other superfamilies, but for most superfamilies, the member is the combination with just one or two domains belonging to other superfamilies. The distribution of the number of combinations also follows a power law[4]. In general, if the superfamily is larger, it may combine with more different types of domains belonging to other superfamilies, according to a power law[6].

How does p-loop containing NTP hydrolases superfamily evolution? How is the distribution of p-loop containing NTP hydrolases superfamily in the three phylogenetic groups? What is the relationship between duplication and combination of p-loop containing NTP hydrolases superfamily? In this paper, we try to answer these questions.

1 Database

We adopt the definition of the domain in structural classification of proteins (SCOP) database developed by Murzin and co-workers[7]. A domain is an evolutionary unit, and it can be duplicated and combined with other domains. The SCOP domains are classified families if they are close evolutionary relatives, usually detectable at the sequence level. Family is brought together into superfamilies which may have low sequence identity but their structural and sometimes functional features strongly suggest a common evolutionary origin[6].

The Superfamily database[8-10] is the source of all domain assignments. It depends on a hidden Markov model homology searching algorithm to search the National Center for Biotechnology Information Entrez Genome database for identification of superfamily fold members[11].

For each genome about half of sequences have been assigned to domains[8].

Following Ref. [11], We take 174 complete genomes and delete five of them (Ddis, Ecun, Pfal, Pyoe and Atum) because their classifications are unsure. The remainder 169 complete genomes are used as the database in this work.

They include 32 eukaryote, 118 bacterial and 19 archaeal genomes. The eukaryotes consist of 17 kind of fungi, 2 kind of plant and 13 kind of animal. We group these genomes into four sets: archaea, bacteria, fungi and plant+animal. For each set we arrange the genomes in order of the amount of contained domains. The archaeal genomes are numbered from 1 to 19, for bacterial genomes from 20 to 137, for fungi from 138 to 154, and for plant and animal from 155 to 169.

2 Abundance Distribution of p-Loop for Archaea, Bacteria, Eukarya

In a genome most of superfamilies contain only one or few domains, but a few superfamilies contain many domains. On the other hand, for different genomes the number of domains belonging to the particular superfamilies is varying because of the extent of duplication of the superfamily. This number of domains is defined as the abundance of the superfamily.

The most common superfamily in archaeal and bacterial genomes is p-loop containing NTP hydrolases superfamily. In eukaryote genomes the most common superfamily is C_2H_2 zinc finger superfamily and the second is p-loop containing NTP hydrolases superfamily. P-loop containing NTP hydrolases superfamily occurs in all 169 complete genomes, but its extent is very different. We calculate the abundance, which is the extent of duplication of p-loop containing NTP hydrolases superfamily in 169 complete genomes shown in Fig. 1(a). Figure 1(a) shows the abundance of p-loop containing NTP hydrolases superfamily increases as the total number of domains in the genome increases. The abundance of p-loop containing NTP hydrolases superfamily in metazoa is much larger than that in the rest. The ratio of the abundance of p-loop to the total number of domains for a genome is defined as the relative abundance of p-loop containing NTP hydrolases superfamily. Figure 1(b) shows the relative abundance of p-loop containing NTP hydrolases superfaily. It can be seen from Fig. 1 (b) that for 169 complete genomes the relative abundance of p-loop containing NTP hydrolases superfamily is very low for metazoa. As an example, for human the number of domains in p-loop containing NTP hydrolases superfamily is a large number of 1 607, but the relative abundance is only 3%.

To sum up, for metazoa although the relative abundance of p-loop is much lower, the number of the domains belonging to the p-loop containing NTP hydrolases superfamily is much higher than those for the rest of 169 complete genomes.

3 Number of Combination Partners in Archaea, Bacteria and Eukarya

In general, if the number of residues between two

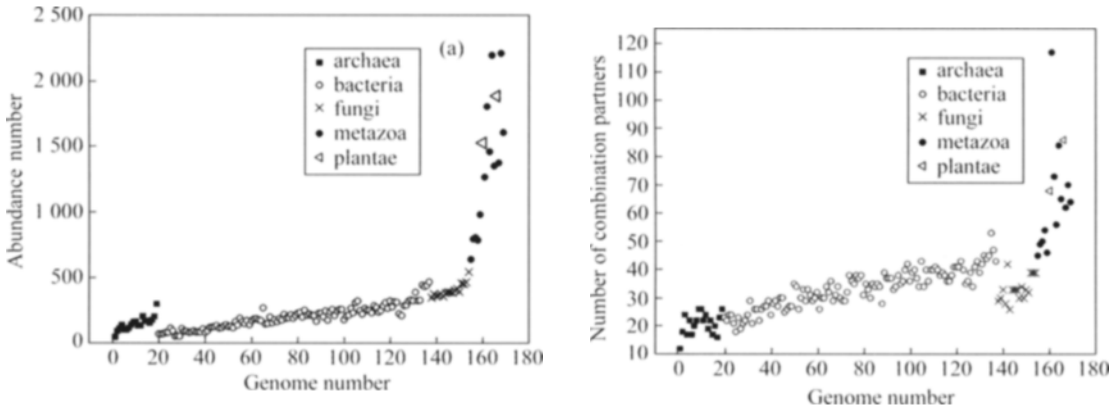

Fig. 1 The abundance (a) and relative abundance (b) distribution of p-loop containing NTP hydrolases superfamily for archaea, bacteria, and eukarya. The eukarya is divided into fungi, plantae and metazoa

neighboring domains in a polypeptide chain is not more than 30, this pair of domains is neighbours and they combine with each other in the course of evolution[6]. The number of combination partners of a given superfamily is the number of superfamily types, whose domains are adjacent to the domains of this given superfamily. The number of combination includes repletion of itself. Figure 2 plots the number of combination partner of p-loop containing NTP hydrolases superfamily for 169 complete genomes. It can be seen from Fig. 2 that the number of combination partners increases slowly from archaea to bacteria, but in eukarya it has a markedly increase. Going into details, for 19 genomes of archaea the number of combination partners, is the adjacent domains of p-loop containing NTP hydrolases superfamily include 66 types of superfamilies, for 118 genomes of bacteria the combination partners of p-loop include 230 types of superfamilies, and for 32 eukarya those include 300 types.

The combination of superfamilies is a very complex problem. The SCOP Release 1. 69 provides 1 539 protein superfamilies. If the combination occurs stochastically, there are potentially 1 $539^2=2\ 368\ 521$ different pairwise combinations. In fact the combinations existed in nature is only a tiny faction of the potential number. Ref. [2] has given an example that in 85 genomes containing two or more domains a total of only 2 500 different pairwise combinations. Ones would expect this to be largely the result of selection for function[2]. The energy for motion and reactions in the cell is often provided by p-loop containing NTP hydrolases. Domains from this superfamily can hydrolyze ATP or GTP and can act as kinases and transferases on their own or combined with different superfamilies[6].

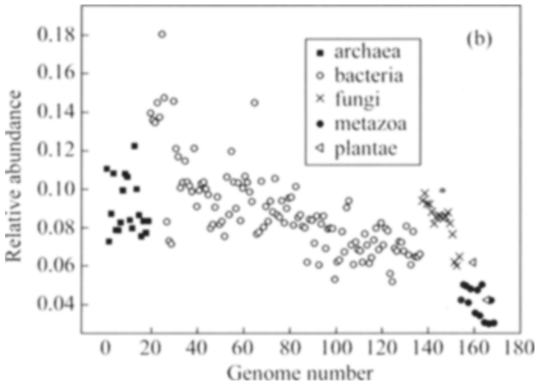

Fig. 2 The number of combination partners for p-loop containing NTP hydrolases superfamily in archaea, bacteria and eukarya. The eukarya is divided into fungi, plante and metazoa

4 Relationship between the Abundance and Number of Combination Partners of P-Loop Containing NTP Hydrolases Superfamily

We compare Fig. 1(a) and Fig. 2, it can be seen the distributions of abundance and combination of p-loop containing NTP hydrolases in 169 completed genomes are resembles. Figure 3 plots the relationship between the abundance and the number of combination partners of p-loop containing NTP hydrolases. In Fig. 3, each dot represents a genome. Although the abundance and the number of combination partners of p-loop containing NTP hydrolases superfamily between metazoa and the rest are very different (see Fig. 1(a) and Fig. (2)), the relationship between the abundance and the number of combination partners of p-loop containing NTP hydrolases follows a power law as $y\text{-}x^{\theta}$ and $\theta \approx 1.92$. In a genome the more the abundance of p-loop containing NTP hydrolases is, the more the number of combination partners is. Therefore, due to the large abundance and number of combination of combination partners there are more function selections for metazoa than those for rest in 169 complete genomes.

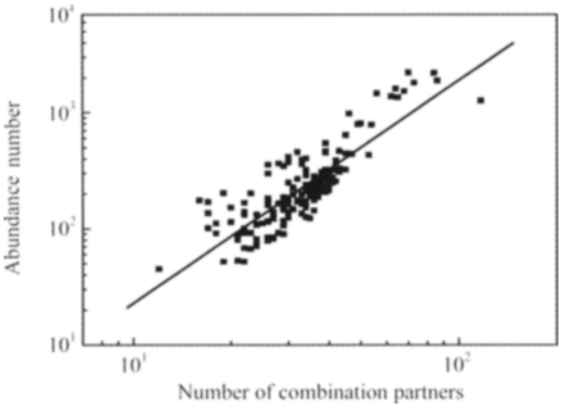

Fig. 3 The relationship between the abundance and number of combination partners of p-loop containing NTP hydrolases. The symbols are the data taken from Superfamily database and the line is the fitting result by y-x^{θ} with $\theta\approx1.92$

5 Conclusion

We provide a quantitative insight into duplication and combinations of p-loop containing NTP hydrolases in archaeal, bacterial, and eukaryote genomes. We analyzed the features of abundance and combination distribution of p-loop containing NTP hydrolases superfamily in 169 complete genomes. P-loop is very common in protein repertoire, and its abundance and the number of combination partners are very large. When the total number of domains is larger in a genome, the abundance and the number of combination partners of p-loop containing NTP hydrolases are more.

Although the abundance and the number of combination of p-loop containing NTP hydrolases between metazoa and the rest is very different, the relationship of abundance and the number of combination partners fits a power-law in archaea, bacteria, and eukarya. P-loop containing NTP hydrolases superfamily combines with more types of different superfamilies when its abundance is larger in a genome.

References

[1] Vogel C, Teichmann S A, Pereira L J. The Relationship between Domain Duplication and Recombination [J]. *J Mol Biol*, 2005,**346**(1):355-365.

[2] Chothia C, Gough J, Vogel C, *et al*. Evolution of the Protein Repertoire [J]. *Science*, 2003,**300**:1701-1703.

[3] Muller A, Maccallum R M, Sternberg M J. Structural Characterization of the Human Proteome [J]. *Genome Res*, 2002, **12**:1625-1641.

[4] Vogel C, Bashton M, Kerrison N D, *et al*. Structure, Function and Evolution of Multidomain Proteins [J]. *Curr Opin Struct Biol*, 2004,**14**:208-216.

[5] Qian J, Luscombe N M, Gerstein M. Protein Family and Fold Occurrence in Genomes: Power-Law Behaviour and Evolutionary Model [J]. *J Mol Biol*, 2001,**313**:673-681.

[6] Apic G, Gough J, Teichmann S A. Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes [J]. *J Mol Biol*, 2001,**310**:311-325.

[7] Murzin A, Brenner S E, Hubbard T, *et al*. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures [J]. *J Mol Biol*, 1995,**247**: 536-540.

[8] Madera M, Vogel C, Kummerfeld S K. The Superfamily Database in 2004: Addition and Improvements [J]. *Nucl Acids Res*, 2004,**32**:D235-D239.

[9] Gough J, Chothia C. Superfamily: HMMs Representing all Proteins of Known Structure. SCOP Sequence Searches, Alignments and Genome Assignments [J]. *Nucleic Acids Res*, 2002,**30**:268-272.

[10] Gough J, Karplus K, Hughey R, *et al*. Assignment of Homology to Genome Sequences Using a Library of Hidden Markov Models that Represent All Proteins of Known Structure [J]. *J Mol Biol*, 2001,**313**:903-919.

[11] Yang S, Doolittle R F, Bourne P. Phylogeny Determined by Protein Domain Content [J]. *Proc Natl Acad Sci*, 2005,**102**(2):373-378.

□