Analysis of Domain Combinations in Eukaryotic Genomes

Masumi Itoh
itoh@kuicr.kyoto-u.ac.jp

Akiyasu C. Yoshizawa
acyshzw@kuicr.kyoto-u.ac.jp

Shujiro Okuda
okuda@kuicr.kyoto-u.ac.jp

Susumu Goto
goto@kuicr.kyoto-u.ac.jp

Minoru Kanehisa
kanehisa@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho,
Uji city, Kyoto 611-0011, Japan

Keywords: eukaryotic genome, domain fusion, comparative genomics

1 Introduction

Domains are basic building blocks of proteins and they can be the units of function, structure and evolution. Proteins often have more than two domains and the combination of domains contribute to achieving a broad functional spectrum and effective evolution of proteins.

Especially for eukaryotes, higher biological mechanisms (for example, cell-cell interactions, processing of information from environment and developmental regulation) are achieved by various domain combinations. Some previous works suggest that there are species- or group- specific domain combinations and that they could be related to specific biological functions in the species [1].

Such comparative genomic analyses need to use whole genome data as much as possible. Here, we collected whole protein sets from completely sequenced and semi-completely sequenced genomes including draft eukaryotic genomes, and we analyzed the domain combinations to obtain an overview of eukaryotic genomes.

2 Materials and Method

The protein sequences of all completely sequenced bacteria and archaea, as well as *Arabidopsis thaliana*, *Caenorhabitis elegans*, *Drosophila melanogaster*, *Encephalitozoon cuniculi*, *Plasmodium falciparum*, *Saccharomyces cerevisiae* and *Schizosaccaromyces pombe* were obtained from KEGG/GENES [4] . We also obtained all predicted protein sequences in draft genomes of *Anopheles gambiae*, *Ciona intestialis*, *Danio rerio*, *Homo sapiens* , *Mus musculus*, *Rattus norvegicus*, *Neurospora crassa*, *Oryza sativa* japonica *Nipponbare*, *Plasmodium yoelii*, *Takifugu rubripes*, *Caenorhabditis briggsae* from Ensembl, TIGR and JGI.

To obtain the domain or motif information of these proteins, we assign Pfam [2] domains using the HMMER programs [3]. For those organisms stored in KEGG/GENES, we used the pre-calculated Pfam domain assignments stored in KEGG/SSDB. For the draft genomes, we performed an HMM search with the same conditions to that used in KEGG/SSDB.

3 Results and Discussion

Table 1 shows the domains which are frequently ovserved in the domain combinations in each species group. The number in parentheses shows the number of unique combinations containing the domain,

Table 1: Most frequent domains observed in each group.

Animal		Fungi	Plant	Protist	Archaea	Bacteria
Deuterostomia	Proteostomia					
pkinase(85)	pkinase(52)	pkinase(19)	pkinase(61)	CPSase(9)	HATPase_c(8)	HATPase_c(13)
EGF(70)	PH(46)	PH(18)	rve(48)	HATPase_c(9)	biotin_lipoyl(7)	CPSase(10)
PH(64)	ank(36)	CPSase(16)	rvt(48)	helicase_C(8)	PAC(6)	biotin_lypoyl(10)
ig(60)	SH3(36)	GATase(14)	zf-CCHC(40)	GATase(7)	ACT(6)	response_reg(9)
zf-C3HC4(55)	EGF(33)	HATPase(12)	PPR(30)	DnaJ(6)	CPSase(6)	Biotin_carb_C(8)
2314	1418	664	1083	233	185	333 (total)

which also means the number of different combination partners. The bottomline shows the total number of domain combinations that we observed in each organism group.

In all species in Crown organisms, the protein kinase (pkinase) domain has the largest number of partners. It is a well known feature that the protein kinase domains have important functions in eukaryotic signaling pathways and it is consistent with previous work [1]. However, other frequently ovserved domains in the combinations are very different between the Animal + Fungi group and the Plant group. These results suggest that the protein kinase domain has key functions in both organism groups, but the methods to use them might be different because domain partners are all different. Anyway, the domain fusion and/shuffle could be an important mechanism. It is also interesting that Protists don't have so many eukaryotic specific domain combinations.

The Pleckstrin homology domain (PH) also participates in the domain combinations in Animal and Fungi. The domain is observed in many proteins in the intracellular signaling pathways. It suggests that at the branching point of the Animals and Fungi, the system including PH and protein kinase domains already existed. This information should be useful to analyze the evolution of regulatory networks.

Acknowledgments

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Apic, G., Gough, J., and Teichmann, S.A., Domain combinaitons in archaeal, eubacterial and eukaryotic proteomes, *J. Mol. Biol.*, 310:311–325, 2001.

[2] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L., The Pfam protein families database, *Nuecleic Acid Res.*, 30:276–280, 2002.

[3] Eddy, S. R., Profile hidden Markov models, *Bioinformatics*, 14:755–763, 1998.

[4] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A., The KEGG databases at GenomeNet, *Nuecleic Acid Res.*, 30:42–46, 2002.