Genomics & Informatics Vol. 7(2) 131-135, June 2009

Comparison of Exon-boundary Old and Young Domains during Metazoan Evolution

Byungwook Lee*

Korean BioInformation Center, KRIBB, Daejeon 305-806, Korea

Abstract

Domains are the building blocks of proteins. Exon shuffling is an important mechanism accounting for combination of a limited repertoire of protein domains in the evolution of multicellular species. A relative excess of domains encoded by symmetric exons in metazoan phyla has been presented as evidence of exon shuffling, and symmetric domains can be divided into old and new domains by determining the ages of the domains. In this report, we compare the spread, versatility, and subcellular localization of old and new domains by analyzing eight metazoan genomes and their respective annotated proteomes. We found that new domains have been expanding as multicellular organisms evolved, and this expansion was principally because of increases in class 1-1 domains amongst several classes of domain families. We also found that younger domains have been expanding in membranes and secreted proteins along with multi-cellular organism evolution. In contrast, old domains are located mainly in nuclear and cytoplasmic proteins. We conclude that the increasing mobility and versatility of new domains, in contrast to old domains, plays a significant role in metazoan evolution, facilitating the creation of secreted and transmembrane multidomain proteins unique to metazoa.

Keywords: domain mobility and versatility, exon shuffling, old and young domains

Introduction

Domains are the building blocks of proteins. Domains functions of multi-domain proteins contribute to our understanding of the proteins (Lee & Lee, 2008). Domains can occur as single-domain proteins or in combinations to form multi-domain proteins (Han, *et al.*, 2007). The analyses of complete genome sequences have revealed that multi-domain proteins have increased in number during evolution; two-thirds of prokaryote proteins contain more than two domains, whereas in eukaryotes about four-fifths of proteins are multi-domain (Chothia, *et al.*, 2003). These data indicate that eukaryote proteins are developing more complex domain architectures and properties, formed by domain duplication and combinations of a limited repertoire of domain types (Ye & Godzik, 2004).

The eukaryote exon-intron structure suggests that domain accretion can be accomplished by the acquisition of exons encoding one or two domains (Liu, *et al.*, 2005). In other words, protein domains tend to be encoded by one exon, or small combination of exons, that begin and end in the splice frame. The duplication and rearrangement of such exons can create novel genes with revised functional properties. This domain reuse has been described as the exon shuffling theory (Patthy, 1999), mediated by intronic recombination of exons encoding protein domains (Eickbush, 1999) and by the action of retrotransposons (Moran, *et al.*, 1999). Proteins that are composed of a number of discrete domains are termed mosaic proteins, and are particularly abundant in metazoan phyla (Kolkman and Stemmer, 2001).

Exon shuffling suggests that the two flanking introns of a domain-encoding exon should have symmetric phase combinations (Patthy, 1999). Intron phases are determined by the examination of the translational reading frame relative to the intron, and introns are thus described as phase 0, 1, or 2 introns. Of the nine possible combinations of flanking introns, three are symmetric (0-0, 1-1, and 2-2) and six are asymmetric. The length of a symmetric exon is always a multiple of three nucleotides. Only symmetric exons can be duplicated in tandem or deleted without affecting the reading frame, whereas the duplication and deletion of asymmetric exons can disrupt the downstream reading frame (Patthy, 1996). Comparisons between human exon-boundary domains and bacterial domains have indicated that 1-1 domains are associated with the origin of animal multicellularity, whereas 0-0 domains are shared between eukaryotes and prokaryotes (Kaessmann, *et al.*, 2002). This comparison also indicates that 0-0 domains date back to before the prokaryote/eukaryote divergence and can thus be defined as 'old domains', whereas 1-1 domains were created recently and can thus be termed 'young domains'.

In this study, we analyzed eight metazoan genomes

*Corresponding author: E-mail bulee@kribb.re.kr
Tel +82-42-879-8531, Fax +82-42-879-8519
Accepted 16 May 2009

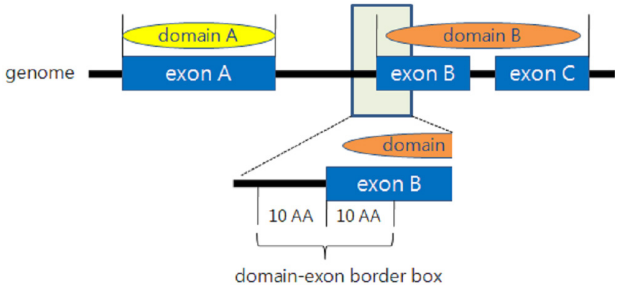

Fig. 1. Illustration of domain and exons. A domain A is encoded by exon A, which is an example of one domain one exon. A domain B is encoded by exons B and C, which is an example of one domain many exons.

and their respective annotated proteomes to explore differences in the spread, versatility, and sub-cellular localization of old and young domains during metazoan evolution.

Methods

Data preparation

We downloaded gene structures and protein sequences of primates (human and chimpanzee), rodents (mouse and rat), dog, fish (zebrafish and tetraodon), and worm (*Caenorhabditis elegans*) from the Ensembl (Hubbard, *et al.*, 2007) database (ftp://ftp.ensembl.org/pub/release45/). The gene data included genomic exon positions, intron phase information, and gene descriptions. Exon phases were obtained by the analysis of the two flanking introns. If several transcripts for any particular gene had been described, we selected the longest coding structure for analysis, to retain the maximum number of domains.

The domain content of protein sequences was analyzed with Pfam 23 (Finn, *et al.*, 2006). In this work, we used Pfam domains rather than structurally defined SCOP (Andreeva, *et al.*, 2004) domains, because Pfam domains offer better coverage of genomes, especially for membrane proteins. The positions in proteins of domain hits, with the cutoff E-value of 0.01, were obtained, and the positions in amino acid coordinates were converted to positions in cDNA sequences. We included only genes with more than one Pfam domains for further analysis.

Identification of exon-boundary domains

Exon-boundary domains in the eight metazoan species were obtained from the positions of cDNA sequences corresponding to Pfam domains, and the positions of exon boundaries. We defined the "domain-exon border box" to determine relationships between domain boundaries and exon boundaries. To be classified as an exon-boundary domain, exon borders were required to be located inside the border boxes [−10, +10] at both ends of a domain (Fig. 1). In forming correlations between a domain and the number of encoding exons, there are three possible relationships: one domain one exon, one domain many exons, and many domains one exon. In this study, we focused on the first two relationships.

Sub-cellular localizations of old and young domains

Sub-cellular localizations of human, mouse, and rat proteins were extracted from the "subcellular location" comments in the UniProtKB/Swiss-Prot database (Wu, *et al.*, 2006). Only the exact and complete matches to one of the following phrases were included: "Nucleus", "Cytoplasm", "Secreted", "Type I membrane protein", or "Type II membrane protein". We excluded proteins with multiple matches and multi-pass membrane proteins. For membrane proteins, protein sequences were divided into cytoplasmic, transmembrane, and extracellular segments.

Results

Identifying the reuse of old and young domains during metazoan evolution

We extracted exon-boundary domains of the eight aforementioned organisms from Ensembl and Pfam results if both ends of the domains encoding exon(s) were located within a "domain-exon border box". Then, exon-boundary domains of each organism were divided into nine classes according to the combinations of their surrounding intron phases. We defined a protein domain flanked by phase 1-1 introns as a class 1-1 domain and a protein domain flanked by phase 0-0 introns as a class 0-0 domain. The numbers of the exon-boundary and the class of domains are given in Table 1. To examine which class domains were expanded during metazoan evolution, we calculated the relative frequency of the nine classes of each species and compared these figures. Each relative frequency was obtained by dividing the number of domains of each class by the total number of exon-boundary domains. The relative frequency of each class in the eight organisms is shown in Fig. 2, which makes clear that the relative frequency of class 0-0 domains is higher than that of class 1-1

Table 1. Summary of exon-boundary domains in eight metazoan species

Species	Proteins[a]	Domains[b]	0-0[c]	1-1[d]	Other comb.[e]
human	16,514	8,363	1,868	3,336	3,402
chimpanzee	15,038	7,993	1,949	2,884	3,381
mouse	18,168	8,358	2,118	2,840	3,553
rat	18,265	9,033	2,812	2,609	3,687
dog	15,262	9,299	2,496	2,868	4,027
zebrafish	19,930	11,746	3,819	2,722	5,024
tetraodon	17,174	8,796	3,141	2,104	3,616
C. elegans	12,260	2,510	885	319	1,335

[a]The number of proteins processed for each species.
[b]The number of exon-bordering domains.
[c]The number of domains with a 0-0 combination of flanking intron phases.
[d]The number of domains with a 1-1 combination of flanking intron phases.
[e]The total number of domains excluding class 0-0 and class 1-1 domains.

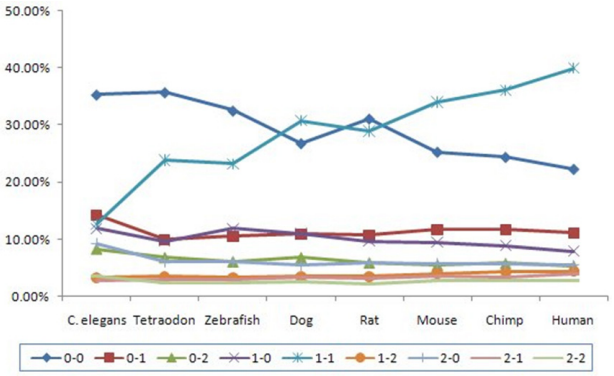

Fig. 2. The relative frequency of nine domain classes in eight organisms. The relative frequency of class 0-0 domains is higher than that of class 1-1 domains in *C. elegans*, tetraodon, and zebrafish, whereas the relative frequency of class 1-1 domains is higher than that of class 0-0 domains in mouse, chimpanzee, and human. The relative frequency of these two domain types was approximately equal in dog and rat.

domains in *C. elegans*, tetraodon, and zebrafish, whereas the relative frequency of class 1-1 domains is higher than that of class 0-0 domains in mouse, chimpanzee, and human. All the remaining classes showed relatively similar frequencies. Because class 0-0 domains are related to old domains and class 1-1 domains are associated with modern domains, as described above, these results indicated that the young domains became widely spread during metazoan evolution whereas old domains became under-represented during this process.

Comparing the versatility of old and young domains

In multi-domain proteins, most protein domains have few partner domains and appear in a highly conserved order (Vogel, *et al*., 2004). Certain domains appear, however, in many unrelated domain architectures. That is, they are mobile and promiscuous domains, characterized by their ability to fold independently. This feature prevents misfolding when such a domain is inserted into a new protein and these domains are typically short and show high versatility (Han, *et al*., 2007) .

We examined partner domains of class 0-0 and class 1-1 domains to identify which might be more versatile. To do this, we extracted N-terminal and C-terminal partner domains from class 0-0 and 1-1 domain of the eight organisms, and then obtained distinct partner domains of the two classes of each domain. The average number of distinct partner domains in the two classes of each domain was calculated by dividing the number of distinct partner domains by the domain frequency. For example, if a domain D appears in 5 proteins and has 10 distinct partner domains, the average number became 2.0 (10/5). This analysis shows that class 1-1 domains had 1.8 partner domain families and class 0-0 domains had 0.7 partner domain families, on average. The other class domains had 0.8-1.0 partner domain families. This indicates that class 1-1 domains are the most versatile, and class 0-0 domains are the most static. Because class 1-1 domains are encoded either by a single exon or by multiple exons, we examined partner domains of both exon types to identify which type is more versatile. Domains encoded by single exons had 2.3 partner domain families and domains encoded by multiple exons had 1.4 partner domain families. This indicated that class 1-1 domains encoded by single exons are the most versatile.

Sub-cellular localization of symmetric domain families

To identify the contributions to multicellularity of old and young domains, we investigated whether there were significant differences in the number of old and young domains in the proteins in different sub-cellular locations. To do this, human protein localization information in the UniProtKB/Swiss-Prot database was used.

From this database, we collected a total of 5,339 human genes where sub-cellular locations were defined as one of the nucleus, the cytoplasm, the extracellular environment (secreted proteins), and the cell membrane.

Table 2. Summary of sub-cellular localization of class 0-0 and 1-1 domains

	Nucleus	Cytoplasm	Extracellular	Membrane		
				Cytoplasmic	Extracellular	Transmembrane
Domains[a]	920	582	1,025	32	922	81
0-0 domains	172	150	72	3	40	1
1-1 domains	87	58	755	1	778	58

[a]The total number of exon-boundary domains.

The sub-cellular localizations of protein domains were determined using the protein sub-cellular locations. Of the 8,363 human exon-boundary domains, the sub-cellular localizations of 3,562 were identified. From these 3,562 domains, class 0-0 and 1-1 domains were selected for study. We observed that the average diversity of class 0-0 domains and class 1-1 domains varied between different sub-cellular compartments (Table 2). For class 0-0 domains, the frequency order was nucleus > cytoplasm > secreted > membrane (p-value < 9.8E-170). On the other hand, the frequency order of class 1-1 domains was membrane > secreted > nucleus > cytoplasm (p-value < 3.7E-289). The domains in membrane proteins were further divided into cytoplasmic, transmembrane, and extracellular domains. Interestingly, most exon-boundary domains in membrane proteins were located in the extracellular region, and most class 1-1 domains of membrane proteins were located in extracellular regions with a small number of domains found in cytoplasmic and transmembrane segments. Class 1-1 domains accounted for most exon-bordering protein domains in extracellular and transmembrane protein segments.

In conclusion, the analyses of sub-cellular localization showed that class 1-1 domains have been extensively reused in the evolution of membrane and secreted proteins, where they are involved in cell-cell signaling, cellular adhesion, and cellular migration, all of which are crucial to the evolution of multicellularity. In contrast, most class 0-0 domains are located in proteins of the nucleus and the cytoplasm, which means that such domains are not directly related to the evolution of multicellularity.

Discussion

The comparative analyses of exon-boundary domains from human and other eukaryotes suggest that young (class 1-1) domains expanded and old (class 0-0) domains contracted during metazoan evolution. This indicates that young domains played important roles in metazoan evolution and the contributions of old domains to multicellularity were relatively small. We found that the expansion of young domains occurred mainly because of expansion of class 1-1 domains amongst the several classes of domain families. The analysis of the versatility of the old and young domains also showed that the young domains are the most versatile; with a versatility index twice that of other domain classes. The analysis of sub-cellular localization indicated that modern domains are mainly located in the proteins of extracellular regions and in extracellular segments of membrane proteins, consistent with previous studies showing that proteins in extracellular regions evolve faster than those of intracellular proteins (Julenius & Pedersen, 2006). Most old domains are located in the nucleus and cytoplasm. In conclusion, the increasing mobility of young domains played a significant role in metazoan evolution, facilitating the creation of secreted and transmembrane multidomain proteins unique to metazoa. In contrast, old domains became less mobile as multicellular organisms evolved.

Acknowledgments

This work was supported by the KRIBB Research Initiative Program and by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MEST) (No. M10869030002-08N6903-00210). BL thanks Jong Bhak for editing and supervising the project.

References

Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., and Murzin, A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* 32, D226-229.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. (2003). Evolution of the protein repertoire. *Science* 300, 1701-1703.

Eickbush, T. (1999). Exon shuffling in retrospect. *Science* 283, 1465-1467.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L., and Bateman, A. (2006). Pfam: clans, web tools and

services. *Nucleic Acids Res.* 34, D247-251.
Han, J.H., Batey, S., Nickson, A.A., Teichmann, S.A., and Clarke, J. (2007). The folding and evolution of multi-domain proteins. *Nat. Rev. Mol. Cell Biol.* 8, 319-330.
Hubbard, T.J., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S.C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A., and Birney, E. (2007). Ensembl 2007. *Nucl. Acids Res.* 35, D610-617.
Julenius, K., and Pedersen, A.G. (2006). Protein evolution is faster outside the cell. *Mol. Biol. Evol.* 23, 2039-2048.
Kaessmann, H., Zollner, S., Nekrutenko, A., and Li, W.H. (2002). Signatures of domain shuffling in the human genome. *Genome Res.* 12, 1642-1650.
Kolkman, J.A., and Stemmer, W.P. (2001). Directed evolution of proteins by exon shuffling. *Nat. Biotechnol.* 19, 423-428.
Lee, B., and Lee, D. (2008). DAhunter: a web-based server that identifies homologous proteins by comparing domain architecture. *Nucleic Acids Res.* 36, W60-64.
Liu, M., Wu, S., Walch, H., and Grigoriev, A. (2005). Exon-domain correlation and its corollaries. *Bioinformatics* 21, 3213-3216.
Moran, J.V., DeBerardinis, R.J., and Kazazian, H.H., Jr. (1999). Exon shuffling by L1 retrotransposition. *Science* 283, 1530-1534.
Patthy, L. (1996). Exon shuffling and other ways of module exchange. *Matrix Biol.* 15, 301-310.
Patthy, L. (1999). Genome evolution and the evolution of exon-shuffling--a review. *Gene* 238, 103-114.
Vogel, C., Berzuini, C., Bashton, M., Gough, J., and Teichmann, S.A. (2004). Supra-domains: evolutionary units larger than single protein domains. *J. Mol. Biol.* 336, 809-823.
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). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucl. Acids Res.* 34, D187-191.
Ye, Y., and Godzik, A. (2004). Comparative analysis of protein domain organization. *Genome Res.* 14, 343-353.