COMPARATIVE PAIR-WISE DOMAIN-COMBINATIONS FOR SCREENING THE CLADE SPECIFIC DOMAIN-ARCHITECTURES IN METAZOAN GENOMES

SHUICHI KAWASHIMA[1] shuichi@hgc.jp
TAKESHI KAWASHIMA[2,3] kawashima38@gmail.com
NICHOLAS H PUTNAM[2] nhputnam@lbl.gov
DANIEL S ROKHSAR[2] DSRokhsar@lbl.gov
HIROSHI WADA[4] hwada@biol.tsukuba.ac.jp
MINORU KANEHISA[1,5] kanehisa@kuicr.kyoto-u.ac.jp

[1] *Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan*
[2] *Department of Molecular and Cell Biology, University of California, Berkeley, LSA Bldg. #3200, Berkeley, CA 94720-3200, USA*
[3] *Japan Sociery for the Promotion of Science, 8 Ichibancho, Chiyoda-ku, Tokyo 102-8472, Japan*
[4] *Graduate School of Life and Environmental Sciences, University of Tsukuba, 1-1-1 Tennoudai, Tsukuba, Ibaraki 305-8572, Japan*
[5] *Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan*

In the evolution of the eukaryotic genome, exon or domain shuffling has produced a variety of proteins. On the assumption that each fusion event between two independent protein-domains occurred only once in the evolution of metazoans, we can roughly estimate when the fusion events were happened. For this purpose, we made phylogenetic profiles of pair-wise domain-combinations of metazoans. The phylogenetic profiles can be expected to reflect the protein evolution of metazoan. Interestingly, the phylogenetic tree of metazoans, derived from the profiles, supported the "Ecdysozoa hypothesis" that is one of the major hypotheses for metazoan evolution. Further, the phylogenetic profiles showed the candidates of genes that were required for each clade-specific features in metazoan evolution. We propose that comparative proteome analysis focusing on pair-wise domain-combinations is a useful strategy for researching the metazoan evolution. Additionally, we found that the extant ecdysozoans share only fourteen domain-combinations in our profiles. Such a small number of ecdysozoan-specific domain-combinations is consistent with the extensive gene-losses through the evolution of ecdysozoans.

Keywords: protein evolution; domain-combination; ecdysozoan; coelomata; phylogenetic tree; metazoa.

1. Introduction

In the evolution of the eukaryotic genomes, exon shuffling or domain shuffling has produced a variety of proteins. Fusion- and fission-events of domains or exons yield new domain-architectures of proteins that may possess novel functions. However, the linkage between such a gain of new domain-architectures and the evolution of metazoans (or animals) remains to be elucidated.

Cloning of the cadherin genes from amphioxus and the finding of its unique domain-architectures provide an opportunity to reconsider the conventional idea of deuterostome phylogeny [15, 16]. Oda and his colleagues found that the domain-architectures of the amphioxus cadherins are similar to invertebrate cadherins rather than tunicate and vertebrate cadherins, although cephalochordata (amphioxus) was long considered as the closest sister group to vertebrates. More recently, two groups reported the close relationship between vertebrate and tunicate by the large amount of sequences from key taxa of metazoans [3, 5].

Considering the case of cadherin, it is expected that comparative analysis of domain-architectures can reveal the evolutional history. Actually, Patthy found that several chordate-specific domain-architectures are related to vertebrate or chordate-specific features [18]. However, it is difficult to compare with the details of domain-architectures for multiple proteomes. In order to address the rapid increase of the draft genomes from variety of metazoan taxa, a convenient method to detect genes including clade-specific domain-architectures is required. Here we propose a new method for comparative domain-architectures, which is useful for comparing among multiple proteomes. Previously, we made the lists of genes with clade-specific domain-architectures [12]. It is noteworthy that the list includes a lot of known genes that have been annotated as their chordate- and vertebrate-specific domain-architectures. In the method, not single domains but pair-wise domain-combinations are regarded as a unit of phylogenetic analysis. The phylogenetic profiles of pair-wise domain-combinations allowed us to survey the timing of the gain and loss of domain-combinations, or gene-fusion and -fission events in the metazoan phylogeny. In the analysis, we found that several vertebrate-specific domain-combinations seem to be required for gaining the vertebrate-specific features such as “cartilage”, “auditory systems” or “tight junction” [12].

Here we extend the analysis of the evolution of the domain architectures to fifteen metazoan genomes and three other higher eukaryotes (fungi, slime mold and plant). Fig. 1 shows the schematic view of the major hypotheses for the phylogeny

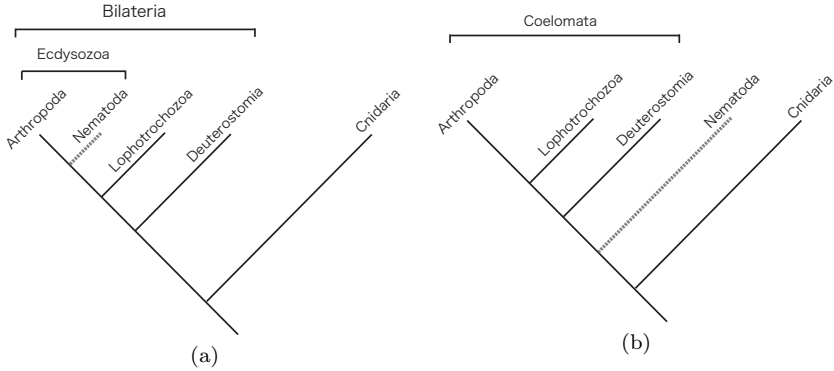

Fig. 1. (a) Ecdysozoa hypothesis. (b) Coelomata hypothesis.

of the metazoa. We highlight particularly the topology of the four branches of the tree, that is Deuterostomia, Arthropoda, Nematoda and Non-bilaterians (including Cnidaria), respectively.

In Fig. 1(a), the nematoda is related to arthropoda in a clade of molting protostomes, termed "Ecdysozoa". Both the Ecdysozoa and Deuterostomia are members of a clade termed "Bilateria" which also includes the third major group termed "Lophotrochozoa" which we do not discuss here. Other non-bilateria groups including Cnidaria are placed at the out-group position of Bilateria. Fig. 1(b) differs from the Fig. 1(a) only in the branch of Nematoda. In Fig. 1(b), Nematoda is sister to the "Coelomata" which includes both Deuterostomia and Arthropoda. Because of their complicated protein domain-architectures, the relationships between protein evolution and its functional innovation is not a simple question, especially in metazoan evolution. Even if there are limited numbers of protein domains in metazoans, their shuffling, fusions and combinations can make a variety of proteins. For example, in the context of two genomes from different species coded the same four protein domains (A, B, C and D), if one species has two genes constructed from "A and B" and "C and D", and meanwhile another species have genes construed from "A and D" and "B and C", can we say that whether they have the same two genes or not? Mutual best-hit search between two species is a simple and convenient method for comparative proteomes, but it may not be powerful as shown in Fig. 2. In Fig 2, a pair of proteins are related by the mutual best-hit. From the view point of domain-combination, however, these are not treated as related genes because no domain-fusion event was shared between these two proteins. Thus, the timing of the domain-fusion event will be revealed by the phylogenetic analysis of domain-architectures. We found there are large differences among the number of clade-specific domain-combinations.

Fig. 2. Schematic view of mutual best hit between two genes containing different domain-architectures.

We can reconstruct a phylogenetic tree from our phylogenetic profiles of pair-wise domain-combinations, when we regard these profiles as a discrete data matrix. We produced such a phylogenetic tree with Phylip Program Package [7] and TreeView [17].

In this phylogenetic tree, two insects and two nematodes are grouped into the same clade. This supports the "Ecdysozoa hypothesis" rather than the "Coelomata hypothesis". Additionally, we found that the very small number of domain-combinations seemed to be Ecdysozoa-specific.

2. Materials and Methods

2.1. *Data*

The following 18 proteomes are used for comparative analysis. Human (*Homo sapiens*): Ensembl build 41 from Ensembl [24], [10], Mouse (*Mus musculus*): Ensembl build 41, Norway Rat (*Rattus norvegicus*): Ensembl build 41, Chicken (*Gallus gallus*): Ensembl build 41, Flog (*Xenopus tropicalis*): JGI v4.1 from JGI [27], Zebra fish (*Danio rerio*): Ensembl build 41, Fugu (*Takifugu rubripes*): JGI version 4.0 from JGI, Ascidian (*Ciona intestinalis*): JGI version 2.0, Amphioxus (*Branchiostoma floridae*): JGI version 1.0, Sea Urchin (*Strongylocentrotus purpuratus*): NCBI gene build 2 version 1 on Baylor's assembly Spur_v2.1 from www.ncbi.nlm.gov, Fly (*Drosophila melanogaster*): Ensembl build 41, Mosquito (*Anopheles gambiae*): Ensembl build 41, Nematodes (*Caenorhabditis elegans*): Wormbase release WS164 from Wormbase [30] and (*Caenorhabditis briggsae*): Wormbase release WS165, Sea anemone (*Nematostella vectensis*): JGI Annotation of Nematostella genome version 1, Fungi (*Saccharomyces cerevisiae*): genome-ftp.stanford.edu saccharomyces_cerevisiae.gff created on july7-2004, Social amoeba (*Dictyostelium discoideum*): dictyBase [23] Full Chromosomes made 10/05/2004 (primary features made 7/11/2005), and Plant (*Arabidopsis thaliana*): TAIR6 updated 11.2005 from NCBI [28].

2.2. *Construction of the pair-wise domain-combinations*

Pfam database (Pfam_ls, release 16.0) was used as the reference domain database [8]. Hammer search for all of the proteomes against Pfam was executed under a threshold set as $1.0e^{-3}$ [26]. Outputs of hmmpfam are parsed by scripts written in Ruby programming language [29] with BioRuby library [22].

The definition of the pair-wise domain-combination is as follows. As a fundamental idea, we enumerate unique pair-wise domain-combinations without taking into account of the order from N-terminus to C-terminus of proteins. For example, when four domains "A", "B", "A" and "C" are lined on a gene in this order, three pair-wise domain-combinations "A and B", "B and C" and "A and C" can be counted. As shown above example, we do not consider the direction of domains on the gene. Therefore, we treat "A and B" and "B and A" as the same combinations. Further, we do not regard the domain repetition "A and A" as pair-wise domain-combinations.

2.3. *Construction of the phylogenetic profiles*

The phylogenetic profiles of domain-combinations of the 18 species are made as follows. For each domain-combination, we count the number of genes having the both Pfam domains which belong to the domain-combination for each organism and put the numbers in the order of the 18 organisms as shown below.

For example, if the numbers of genes coded a domain-combination are 3, 2, 1,

0, 0, 0 in species S_1, S_2, S_3, S_4, S_5, S_6, respectively, the profile is treated as the following:

$$[S_1, S_2, S_3, S_4, S_5, S_6] \mapsto [3, 2, 1, 0, 0, 0]$$

The whole phylogenetic profiles are shown in the supplemental Table S1. In the table, each row represents a phylogenetic profile of a pair-wise Pfam domain-combination. The first and the second column represents two accession numbers of the Pfam domains which belong to the combination, respectively. Each phylogenetic profile is shown as a set of numbers written in columns from the third to the 20th. From the third to the 20th column, the corresponding 18 species are, *H.sapience*, *M.musculus*, *R.norvegicus*, *G.gallus*, *X.tropicalis*, *D.rerio*, *T.rubripes*, *B.floridae*, *C.intestinalis*, *S.purpuratus*, *D.melanogaster*, *A.gambiae*, *C.elegans*, *C.briggsae*, *N.vectensis*, *D.discoideum*, *S.cerevisiae* and *A.thaliana*, respectively.

Phylogenetic profiles of domain-combinations for the super phyla are made as follows. The columns of phylogenetic profiles of the 18 species (in Table S1) are divided into the following three super phyla or an outgroup. Ten species are assigned to Deuterostomia (*H.sapience*, *M.musculus*, *R.norvegicus*, *G.gallus*, *X.tropicalis*, *D.rerio*, *T.rubripes*, *B.floridae*, *C.intestinalis*, *S.purpuratus*), two species to Arthropoda (*D.melanogaster*, *A.gambiae*), two species to Nematoda (*C.elegans*, *C.briggsae*), four species to Out-group (*N.vectensis*, *S.cerevisiae*, *D.discoideum*, *A.thaliana*). The result of phylogenetic profiles of the super phyla is shown in Table 1.

Phylip program [7] was used for making the phylogenetic tree from the phylogenetic profile of pair-wise domain-combinations. We converted the Table S1 into binary data (i.e. 0 to 0 and otherwise to 1) and used it as input data for Pars that is a parsimony program included in the Phylip package. The phylogenetic tree was shown in Fig. 3.

3. Results

In our previous study, we made lists of the several clade-specific genes in deuterostomes. The functional annotations and experiments of the genes strongly suggested links between their functions and their clade-specific features; for example, chordate-specific genes included notochord related genes and vertebrate-specific genes included cartilage specific genes, etc. [12].

These results mean that our method is effective for understanding the relationships between protein evolution and biological features in the evolution of deuterostomes. Here, we would like to focus on another key animal clade, Ecdysozoa, which includes two big sub-clades, Nematoda and Arthropoda. Well-annotated genomes of *C.elegans* and *D.melanogaster* are available as representatives of both clades. We selected the genes containing the ecdysozoan-specific domain-combinations. Interestingly, we found only fourteen genes satisfying the above condition and seven of the fourteen were common to the ecdysozoans (Table 1 "_AN_" and Table 2).

Table 1. The numbers of domain-combinations shared among four super phyla.

The type of profiles[a]	The number of shared domain-combinations[b]	The number of shared class-I domain-combination[c]	The number of shared class-II domain-combination[d]
D A N O	1,468	0	1,468
D A N _	298	22	276
D A _ O	372	0	372
D _ N O	120	0	120
_ A N O	0	0	0
D A _ _	236	21	215
D _ N _	106	4	102
_ A N _	18	4	14
D _ _ O	572	0	572
_ A _ O	9	0	9
_ _ N O	10	0	10
D _ _ _	3,702	236	3,466
_ A _ _	100	12	88
_ _ N _	176	17	159
_ _ _ O	1,297	0	1,297
Total	8,484	316	8,168

Note: [a] D: Deuterostomia (*H.sapience*, *M.musculus*, *R.norvegicus*, *G.gallus*, *D.rerio*, *T.rubripes*, *B.floridae*, *C.intestinalis* and *S.purpuratus*). A: Arthropoda (*D.melanogaster* and *A.gambiae*). N: Nematoda (*C.elegans* and *C.briggsae*). O: Out-group of bilaterians (Cnidaria (*N.vectensis*), and other higher eukaryotes (*S.cerevisiae*, *D.discoidium* and *A.thaliana*)). Underscore ("_") means that the corresponding organismal-group do not have the domain-combinations. (E.g., the row of "DAN_" represents the number of domain-combinations that are shared between Deuterostomia, Arthropoda and Nematoda but not with out-group of bilaterians). [b] The total of the third column and fourth column is shown in the second column. [c] The class-I domain-combination defined as the each individual domains is not found in Out-group. [d] The class-II domain-combination defined as the both of the individual domains are found in Out-group.

Thus, the number of domain-combinations gained in the ecdysozoans is quite small, compared to that gained in deuterostomia (3466 in Table 1, "D _ _ _"). We examined the meaning of differences between these two numbers blow.

We constructed a phylogenetic tree from the phylogenetic profiles. (Fig. 3). In the tree, the two species of Arthropoda (*D.melaogaster* and *A.gambiae*) and the two species of Nematoda (*C.elegans* and *C.briggsae*) form a clade. This means that information from gain and loss of domain-combinations supports the "Ecdysozoa hypothesis" rather than the "Coelomata hypothesis".

Table 1 shows the phylogenetic profile of the domain-combinations from three clades of super phyla (Arthropoda, Nematoda and Deuterostomia) and the out-groups (cnidarian, fungi and plant). Since all combinations in the fourth column of Table 1 are of two modules both of which are found in out-group species, it is expected that these numbers represent the gains of combinations from pre-existing domains. Because all of the single domains counted on the fourth column in Table 1 are found in out-group species, it is expected that these numbers represent the gains of

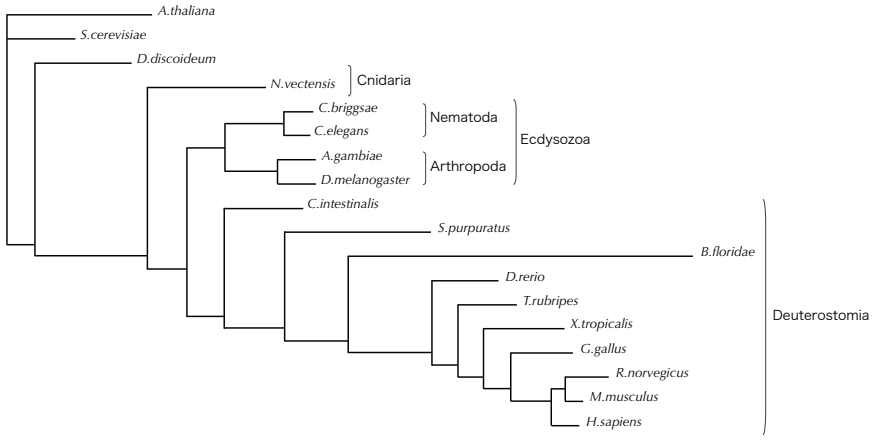

Fig. 3. Parsimonious tree based on domain-combinations from 18 species.

domain-combinations from pre-existing two individual domains. In the Table 1, the number of common domain-combinations within Deuterostomia and Arthropoda is 215. In contrast, the number of common domain-combinations within Arthropoda and Nematoda is only 14. Thus, if we ascribed the gain of domain-combinations to the important factors which influence the topology of the phylogenetic tree, these two numbers may support the "Coelomata hypothesis".

However, the counterpart that is the loss of domain-combinations also influences the topology. Therefore the topology of branches supports the "Ecdysozoa hypothesis". The number of loss of domain-combinations which is common to Arthropoda and Nematoda is 572 from out-group, and the corresponding number to Deuterostomia and Nematoda is merely 10. This result supports a scenario in which extensive gene losses have occurred in the evolution of Ecdysozoans. This data is consistent with the result from comparative proteomes between human and cnidarians [13, 14, 19, 20]. EST analyses of Cnidaria that is one of the animal clades as out-group of bilaterians (Fig. 1) revealed that a substantial number of the ESTs appeared to be vertebrate homologs but no counterpart in the fly or worm.

Interestingly, Nematostella is located on the basal position of metazoans in Fig. 3, even though the number of mutual best-hit genes between human and nematostella is larger than those between others. This result suggests an advantage of our method which can distinguish the two sequences of similar proteins that have different domain-architectures. The best-hit search between human and Nematostella may possibly include the best-hit single domains which are partly constituting more complicated domain-architectures (Fig. 2). In fact, 3,271 domain-combinations are found only in Deuterostomia, although both of the single domains for the all of domain-combinations are found in an out-group (column 4 in Table 2). The 3,271 domain-combinations may be critical to the evolution of Deuterostomes.

Table 2. The list of the domain-combinations that are ecdysozoan-specific and that are shared between Nematodes (*C.elegans* ans *C.briggsae*) and Arthropods (*D.melanogaster* and *A.gambiae*). The Combinations of the two Pfam-domains ("Pfam accession" of the first column and second column) are found at least one gene in both of Nematodes and Arthropods.

Pfam accession	Pfam accession	Gene name[a]	Gene annotation[a]
PF06421	PF08477	Waw (fly)	Elongation factor-type GTP-binding protein
PF01011	PF07714	PEK (fly)	Unknown
PF00024	PF00100	NompA (fly)	Detection of mechanical stimulus during sensory perception of sound
PF00989	PF07885	Egl-2 (worm)	Voltage-gated potassium channel
PF00520	PF00989	Egl-2 (worm)	Voltage-gated potassium channel
PF00047	PF06582	Ketn-1 (worm)	Actin-bingind protein
PF00014	PF02014	Spon-1 (worm)	Extracellular matrix proteins
PF00014	PF06468	Spon-1 (worm)	Extracellular matrix proteins
PF00795	PF01230	Nft-1 (worm)	Carbon-nitrogen hydrolase
PF01822	PF02485	Sqv-6 (worm)	Peptide O-xylosyltransferase
PF00023	PF05186	F34D10.7 (worm)	Unknown
PF00096	PF01426	MTA1-like (fly)	Interacts with two or more components of the EGF/RAS signalling pathway
PF00041	PF01682	CG17839 (fly)	Unknown
PF00047	PF01682	CG17839 (fly)	Unknown

Note: [a] A representative gene name is shown in the third column ("Gene name"). Each annotation for the gene is shown in the fourth column ("Gene annotation"). Gene names and annotations are quoted from the wormbase [30] and the flybase [25].

4. Discussion

In this research, we employed the pair-wise domain-combinations rather than combination of more larger number of domains. e.g. ternary or quartette. Because we have focused evolutionary events when two different domains fused into one gene, not domain-architectures on genes itself. For another reason, the number of the data will be reduced if we considered more large number of domain combination. In addition, it is not suitable to utilize the detailed domain-architectures of genes in draft genomes for phylogenetic analysis. Meanwhile, even if the gene-models are still draft, it would be rarely seen that the same false-positive domain-combinations are detected throughout all genomes from one clade. Thus, we suggest that pair-wise domain-combinations are useful unit for the comparative analysis of domain-architectures of "draft" invertebrate genomes.

The phylogenetic profile of the pair-wise domain-combinations allows us to reconstruct a phylogenetic tree by a parsimonious method. In the tree, the phylogenetic location of the most of the species are consistent with that of the "Ecdysozoa hypothesis" [1] except *C.intestinalis* (Fig. 3). Here, we focused on the two important points in the hypothesis. One is that the several groups of molting animals, like nematodes, insects and other arthropods, form one clade named as Ecdysozoa. Another, the Cnidarians are the closest animal group of Bilateria (Fig. 1(a)). For the past 10 years, several analyses for the phylogenetic tree of metazoans supported the

"Ecdysozoa" hypothesis but others supported another traditional scenario, "Coelomata hypothesis".

Wang and Caetano-Anolles reported a global phylogeny of 185 organisms, based on the phylogenetic profiles of the domain-architectures among their proteomes [21]. In this respected work, they made profiles from not pair-wise but detail classification of domain-architectures. Interestingly the sub-tree of their phylogeny corresponding to metazoans supports the "Coelomata hypothesis", in contrast with our phylogenetic tree which supports the "Ecdysozoa hypothesis" (Fig. 1(a) and Fig. 3). Our study is the first to support the "Ecdysozoa hypothesis" from the perspective of domain architectures.

Several recent genome analyses of Cnidarians revealed that a significant proportion of genes that are not in ecdysozoa are shared between cnidarians and vertebrates [13, 14, 19, 20]. This suggests that a lot of genes are lost from a variety of phyla in Bilateria after the divergence from the common ancestor of cnidarians. If we stand with the "Ecdysozoa hypothesis", the large (572) number of common-loss of domain-combinations in Insecta and Nematoda is consistent with their monophyletic evolution. The small number (14) of common-gain of domain-combinations suggests that the divergence of Insecta and Nematoda occurred at an early stage after the event when the losses of domain-combinations happened in the common ancestor of ecdysozoan.

Conceivably, large number of gene losses would be independently occurred in the both Arthropoda and Nematoda. Also in this case, loss of genes common in the both species can be secondarily detected. For examining this possibility, we are now analyzing ESTs of another Arthropod to survey their domain-combinations. It would be interesting that what-like domain-combinations can be found in Arthropod other than Insecta.

A tunicate, *C. intestinalis*, locates on the basal position of Deuterostomia in both of the trees in our data as well as in data by Wang and Caetano-Anolles [21]. It seems to be misplacement because the echinoderms have been believed to be branched off earlier than Ciona or Amphioxus [9]. This can be explained by the substantial gene-loss in Tunicate [11]. The Tunicate-specific gene-loss didn't disrupt the position of Deuterostomia in the tree, but it displaces the location of *C.intestinalis* onto the basal position in Deuterostomia.

Thus, we think that the phylogenetic trees made from domain-architectures are not so robust as molecular phylogenetic trees. However, we would like to emphasize the value of the phylogenetic profiles of domain-architectures. We can perceive the past evolutional history from the profiles. The genes common in each phylogenetic node suggest that they were essential genes for evolution of each clade. In Table 2, we listed the ecdysozoan-specific domain-combination genes.

In a recent study, Bashton and Chothia asked how domains affect each other in multi-domain proteins, finding that often they interact in ways that create possibly-associated but new functional attributes [2]. They surveyed the SCOP database and assembled 45 sets of proteins, each containing a multi-domain protein and

homologous 1-domain proteins. In some cases, the combination and interaction of domains may changes the function of a single domain to a related but different one in the context of the more complicated protein. The strategy of our research is similar to the report by Bashton and Chothia. The novelty of our study is that we mapped the domain-combinations on the metazoan-phylogeny. This mapping-step gives us the additional information regarding metazoan evolutions. In the previous study, we proposed a possibility in which the genes containing vertebrate-specific domain-combinations cause several vertebrate innovations, like cartilage formation, tight junction and auditory system [12]. In this report, we found that only fourteen domain-combinations are shared in the four of the extant ecdysozoans (Table 1, column 4, Table 2). Because such global gene-losses seem to be happened in the ancestors of ecdysozoans, the remained fourteen genes are good candidates to study the evolution of the ecdysozoans.

Acknowledgments

We thank Dr. David Pile, Dr. David Hendrix and Dr. Keisuke Kawashima for critical reading of our manuscript. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, 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 and the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo. This work was also supported by JSPS Postdoctral Fellowships for Research Abroad.

References

[1] Aguinaldo, A.M., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff RA, and Lake, J.A., Evidence for a clade of nematodes, arthropods and other moulting animals, *Nature*, 387:489–493, 1997.

[2] Bashton, M., and Chothia, C., The generation of new protein functions by the combination of domains., *Structure*, 15:85-99, 2007.

[3] Bourlat, S.J., Juliusdottir, T., Lowe, C.J., Freeman, R., Aronowicz, J., Kirschner, M., Lander, E.S., Thorndyke, M., Nakano, H., Kohn, A.B., Heyland, A., Moroz, L.L., Copley, R.R. and Telford, M.J., Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida, *Nature*, 444:85–88, 2006.

[4] Cordaux, R., Udit, S., Batzer, M.A., and Feschotte, C., Birth of a chimeric primate gene by capture of the transposase gene from a mobile element, *Proc. Natl. Acad. Sci. U.S.A.*, 103:8101-8106, 2006.

[5] Delsuc, F., Brinkmann, H., Chourrout D., and Philippe, H., Tunicates and not cephalochordates are the closest living relatives of vertebrates, *Nature*, 439:965-968, 2006.

[6] Dopazo, H., and Dopazo, J., Genome-scale evidence of the nematode-arthropod clade, *Genome Biol.*, 6:R41, 2005.

[7] Felsenstein, J., PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2005.

[8] 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., Pfam: clans, web tools and services, *Nucleic Acids Res.*, 34:D247-51, 2006.
[9] Holland, L.Z., A chordate with a difference, *Nature*, 447:153–155, 2007.
[10] 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., et al., Ensembl 2007, *Nucleic Acids Res.*, 35:D610-617, 2006.
[11] Hughes, A.L., and Friedman, R., Loss of ancestral genes in the genomic evolution of Ciona intestinalis, *Evol. Dev.*, 7:196-200, 2005.
[12] Kawashima, T., Kawashima, S., Tanaka, C., Murai, M., Yoneda, M., Putnum, N.H., Rokhser, D.S., Kanehisa, M., Satoh, N., and Wada, H., Module shuffling preformed essential role during the evolution of vertebrates, In preparation.
[13] Kortschak, R.D., Samuel, G., Saint, R., and Miller, D.J., EST analysis of the cnidarian Acropora millepora reveals extensive gene loss and rapid sequence divergence in the model invertebrates, *Curr Biol.*, 13:2190-2195, 2003.
[14] Miller, D.J., Ball, E.E., and Technau, U., Cnidarians and ancestral genetic complexity in the animal kingdom, *Trends Genet.*, 21:536-539, 2005
[15] Oda, H., Wada, H., Tagawa, K., Akiyama-Oda, Y., Satoh, N., Humphreys, T., Zhang, S. and Tsukita, S., A novel amphioxus cadherin that localizes to epithelial adherens junctions has an unusual domain organization with implications for chordate phylogeny, *Evol. Dev.*, 4:426-434, 2002.
[16] Oda. H., Akiyama-Oda, Y., and Zhang, S., Two classic cadherin-related molecules with no cadherin extracellular repeats in the cephalochordate amphioxus: distinct adhesive specificities and possible involvement in the development of multicell- layered structures, *J. Cell Sci.*, 117:2757-2767, 2003.
[17] Page, R.D., TreeView: an application to display phylogenetic trees on personal computers, *Comput. Appl. Biosci.*, 12:357-358, 1996.
[18] Patthy, L., Modular assemlby of genes and the evolution of new functions, *Genetica*, 118:217-231, 2003.
[19] Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V., Jurka, J., Genikhovich, G., Grigoriev, I.V., Lucas, S.M., Steele, R.E., Finnerty, J.R., Technau, U., Martindale, M.Q., and Rokhsar, D.S., Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization, *Science*, 317:86-94, 2007.
[20] Technau, U., Rudd, S., Maxwell, P., Gordon, P.M., Saina, M., Grasso, L.C., Hayward, D.C., Sensen, C.W., Saint, R., Holstein, T.W., Ball, E.E., and Miller, D.J., Maintenance of ancestral complexity and non-metazoan genes in two basal cnidarians, *Trends Genet.*, 21:633-639., 2005.
[21] Wang, M., and Gaetano-Anolles, G., Global Phylogeny Determined by the Combination of Protein Domains in Proteomes, *Mol. Biol. Evol.*, 23:2444-2454, 2006.
[22] `http://bioruby.net/`
[23] `http://dictybase.org/`
[24] `http://www.ensembl.org/`
[25] `http://flybase.bio.indiana.edu/`
[26] `http://hmmer.janelia.org/`
[27] `http://www.jgi.doe.gov/`
[28] `http://ncbi.nlm.nih.gov/`
[29] `http://ruby-lang.org/`
[30] `http://www.wormbase.org/`