The evolution of protein domain families

Marija Buljan and Alex Bateman[1]
Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, U.K.

Abstract

Protein domains are the common currency of protein structure and function. Over 10 000 such protein families have now been collected in the Pfam database. Using these data along with animal gene phylogenies from TreeFam allowed us to investigate the gain and loss of protein domains. Most gains and losses of domains occur at protein termini. We show that the nature of changes is similar after speciation or duplication events. However, changes in domain architecture happen at a higher frequency after gene duplication. We suggest that the bias towards protein termini is largely because insertion and deletion of domains at most positions in a protein are likely to disrupt the structure of existing domains. We can also use Pfam to trace the evolution of specific families. For example, the immunoglobulin superfamily can be traced over 500 million years during its expansion into one of the largest families in the human genome. It can be shown that this protein family has its origins in basic animals such as the poriferan sponges where it is found in cell-surface-receptor proteins. We can trace how the structure and sequence of this family diverged during vertebrate evolution into constant and variable domains that are found in the antibodies of our immune system as well as in neural and muscle proteins.

Introduction

Protein domains are compact regions of a protein's structure that often convey some distinct function. Domain architecture, or order of domains in a protein, is frequently considered as a fundamental level of protein functional complexity [1]. The majority of the protein repertoire is composed of multidomain proteins; two-thirds of the proteins in prokaryotes and about four-fifths eukaryotic ones have two or more domains [2]. Moreover, an organism's complexity relates much better to the number of distinct domain architectures [3] and expansion in particular domain families [4] than to the number of genes in the organism. The prevalence of proteins with more than two domains and the recurrent appearance of the same domain in non-homologues proteins show that functional domains are reused when creating new proteins. Because of this, domains have been likened to Lego bricks that can be recombined in various ways to build proteins with completely new functions [5]. Hence, one way to study evolution of protein function and structure is by looking at the evolution of protein domain composition. The average length of a protein domain is approx. 120 amino acids, so changes in domain architecture are underlined by large alterations at the gene level. Examples of molecular mechanisms that can direct these rearrangements are gene fusion and fission [6], exon shuffling through intronic recombination [7], alternative gene splicing and retropositioning [3]. However, although there is evidence that, in prokaryotes, changes in protein domain composition are directed by gene fusion and fission [8,9], little is still known about exact mechanisms that underlie these changes in eukaryotes [3,6]. Apart from being dependent on the mechanisms that determine them, existing domain combinations are also the result of selective forces that enable them to remain in a population. Interestingly, some domains are observed in a number of different domain combinations, and are considered to be 'promiscuous', whereas others occur in only one or a few combinations [10]. These promiscuous domains are, typically, involved in protein–protein interactions, and some of them play important roles in signalling pathways [11]. This, together with the fact that they show evidence of strong purifying selection affecting them [11], implies that these domains were able to become promiscuous in the first place because they had a potential to be useful in various contexts.

Although, in general, the function of a protein domain is conserved over time, it may be altered or completely changed. A domain sequence that has diverged by mutations, deletions and insertions, if selected, eventually becomes a new different domain whose structure and/or function varies from the original one [12]. Nonetheless, even few subtle changes, such as point mutations, can have dramatic effects on a function of a protein domain and consequently affect the overall protein function. For example, amino acids in an enzyme's active site are usually highly conserved, and their mutations can completely destroy the original function. However, it was found that substitutions of the active-site residues can lead to catalytically inactive forms that can later adopt new functions, such as those in regulatory processes [13]. Additionally, mutations in an enzyme's catalytic site can adapt its specificity to a different substrate, and there are examples of enzymes that have evolved to catalyse different reactions on the same structural scaffold using that mechanism [14]. As evident from the example of enzymatic domains, different residues in proteins evolve at different rates. Similarly, different proteins

Key words: domain shuffling, immunoglobulin superfamily, multidomain protein, protein domain, protein evolution.

Abbreviations used: C-set, constant; I-set, intermediate; V-set, variable.

[1] **To whom correspondence should be addressed (email agb@sanger.ac.uk).**

as a whole also differ in rates of evolution. Some families such as histone proteins change sequence at a very low rate [15], whereas molecules of the immune system such as antibodies change at a 100-fold higher rate [16].

Evolution at protein termini

In previous work where protein evolution has been studied from the domain perspective, homology was assumed between the proteins with similar domain architectures, and differences in domain composition were looked for. One of the significant conclusions was that changes in domain architecture preferentially occur at protein termini [17,18]. Weiner et al. [18] have suggested that all observed changes can be approximated as domain deletions since they are expected to be more frequent than domain insertions. A proposed explanation was that the position of domain gain and loss is defined by the dominant causative molecular mechanisms, which would then be those acting at the termini: gene fusion and fission and in particular insertions of new start and stop codons [18]. We have addressed the same question of domain architecture evolution by using phylogenies from the animal gene families in the TreeFam database (release 4) [19]. For domain assignments, we have used the Pfam [20] protein domain annotation (release 22). The Pfam database is a large collection of protein domains and families with over 10000 families collected so far [21]. The main advantages of this approach compared with the previous studies are that: first, one can be more certain that the proteins being compared are true homologues; secondly, domain composition of the ancestral sequences can be inferred and in that way domain gains differentiated from domain losses; and finally, Treefam allows us to distinguish gene duplication from gene speciation events.

To infer domain composition of ancestral proteins, we have applied the maximum parsimony algorithm [22], which finds the evolutionary scenario that is explained with the fewest gain or loss events. Our results corroborate the previous findings with regard to the preferred gain and loss of domains at protein termini (Figure 1). Moreover, we show that the same distribution of changes, where they tend to occur at the termini rather than in the middle of proteins, goes for both domain gains and domain losses. However, rather than explaining the pattern solely by the causative mechanisms, we believe that the observed distribution of changes is a consequence of the interplay of both mechanisms acting to add and remove domains from protein termini, and also selective forces that disfavour gains and losses of domains within a protein. Protein termini are normally charged, flexible and found at the surface of proteins, so it is easy to imagine that additions or deletions of domains there are less likely to disrupt the rest of the structure, especially if the concerned domains are independent structural units. On the other hand, connector regions between domains direct the contact and interaction of domains they link together. Hence, even if those regions themselves are unstructured and do not have a functional role, it is still more likely that changes there will disrupt the rest of the structure. Therefore it is to be expected that natural selection will prefer changes at the termini to those that occur between the existing domains.

In addition, comparison of preferred positions of changes that were preceded by gene speciation to those preceded by gene duplication does not show significant difference between these two types of event. The latter implies that the same basic mechanisms and evolutionary forces shape emergences of new domain architectures after both types of evolutionary event. However, the frequency with which domain gains and losses are observed is almost 2-fold greater after gene duplication (Table 1). That observation can be explained by the fact that it is more permissive to experiment with new domain composition when the gene exists in two copies [23].

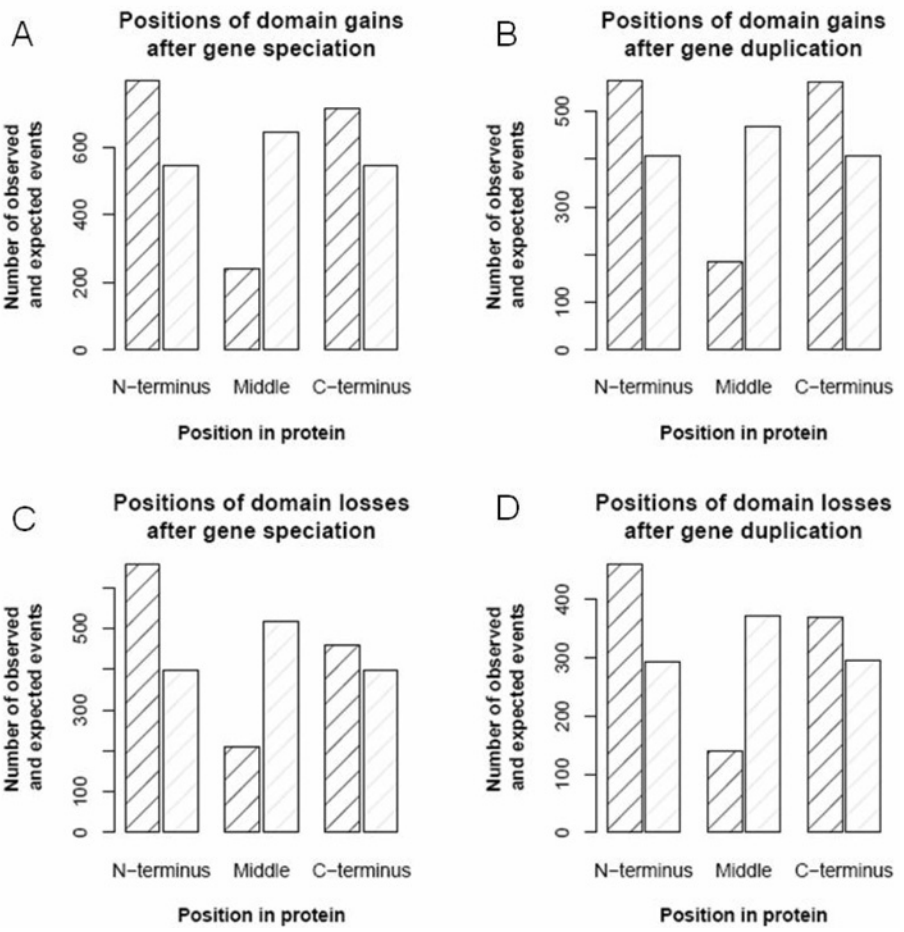
Domains that we have observed to be most frequently gained during animal gene evolution either have a role in extracellular processes or in cell regulation such as signal transduction or DNA binding. Examples of domains in the former category are the EGF (epidermal growth factor) domain, the immunoglobulin superfamily and the CUB (complement protein subcomponents C1r/C1s, urchin embryonic growth factor and bone morphogenetic protein 1) domain, and those in the latter category are zinc finger (C2H2 type), leucine-rich repeat, SH3 (Src homology 3) domain, the PH (pleckstrin homology) domain and RING (really interesting new gene)-finger superfamily. For the same domains (apart from the leucine-rich repeat protein family), Vogel and Chothia [4] have found previously that the number of genes with those domains has a strong correlation with organism complexity. Hence, it was suggested that they could be responsible for the emergence of new complex traits in metazoans. Vogel and Chothia [4] have assigned expansion of these domains primarily to duplications of the genes that have already contained them. However, our results suggest that insertion of these domains into genes that have not previously coded for them has also contributed to their expansion.

Evolution of the structure of the immunoglobulin superfamily

The immunoglobulin superfamily is one of the largest and most diverse groups of proteins in the human genome. It therefore provides an interesting set of proteins to explain protein evolution at longer timescales. The immunoglobulin domain was first identified as a region of sequence similarity of approx. 100 amino acids in length that was found repeated in the antibodies. Later structural work on antibodies showed that they possess a β-sandwich structure with seven to nine strands. The structures also defined that they contained variable or V-set domains and constant or C1-set domains, which differed in their number of β-strands. Later, further structurally distinct sets called the C2-set and intermediate I-set were identified [24]. It had been suggested that the V-set domains or C2-set domains are most likely to be the primordial immunoglobulin-like domain. Looking at the taxonomic range of each structural set allows us to see which, if any, of the structural sets is the ancestral one. The results of this analysis are shown in Table 2.

Figure 1 | Positions in proteins where domain gains (A and B) and losses (C and D) have been observed after gene speciation (A and C) and duplication (B and D)

Observed numbers of events are presented as dark-hatched columns. Expected numbers of gains and losses (light-hatched columns) were calculated based on the representation of ancestral proteins as strings of domains and an assumption that it is equally probable to observe a gain or loss of a domain on any position in that string. There are more events in total following the speciation of genes. However, there are also far more speciation nodes in general in the TreeFam trees (release 4). The results presented do not include gains and losses of repeated domains. The results are obtained after parsing all TreeFam trees with the maximum parsimony algorithm. The bias for the changes to occur at the termini is evident in all categories of events.



The species distribution of immunoglobulin domains shows that the C1-set are only found in vertebrates and therefore evolved late in metazoan evolution. Presently, only I-set and V-set domains are found in sub-vertebrate species. From these results, we can begin to trace the evolution of the structure of immunoglobulin-like domains. The C2-set, that includes proteins such as CD2, CD4 and other cell-surface receptors, evolved in the protostome lineage. The most basal genome presented here is the starlet sea anenome *Nematostella vectensis* which contains both I-set and V-set domains. On the basis of this analysis, we are still not able to distinguish which is the most ancestral set of immunoglobulin domains. Immunoglobulin-like domains have been found in the earliest metazoa, such as poriferans (marine sponges) [25]. Many of these immunoglobulin-like domains have been found associated with kinase domains, suggesting that the immunoglobulins' ancestral function involved signalling. Complete genome sequences from these lower metazoa will give us clues about the structure and functions of the primordial immunoglobulin-like domains.

Conclusions

Protein evolution is evident at different scales of events. On the small scale, single amino acids are mutated, and, on the large scale, whole domains are lost or gained in the protein. An example of a protein family that has experienced a large amount of changes in different lineages is the immunoglobulin superfamily. Members of this superfamily play an essential role in the vertebrate immune response,

Table 1 | Frequency of the observed changes (domain gains and losses) following speciation compared with following duplication of genes

Frequency is stated as both the average number of events (or nodes in the trees) and averge branch length in the trees after which a change is observed. All TreeFam families were taken into account in calculations.

Evolutionary event	Average number of events after which the change is observed	Average branch length after which the change is observed
Gene duplication	36	4.88
Gene speciation	67	8.49
Ratio of speciation over duplication	1.83	1.74

Table 2 | Taxonomic distribution of immunoglobulin superfamily structural sets as defined by the Pfam database

Numbers in the Table refer to the number of sequences that possess a domain in each structural set.

Taxonomic group	Classification	V-set	I-set	C1-set	C2-set
Human	Deuterostome	951	220	388	47
Mouse	Deuterostome	764	190	269	50
Zebrafish	Deuterostome	274	128	88	20
Drosophila melanogaster	Protostome	64	138	0	26
Caenorhabditis elegans	Protostome	19	58	0	0
Sea anenome	Cnidarian	18	101	0	0
Saccharomyces cerevisiae	Fungus	0	0	0	0

cell adhesion and many other processes [26]. They are also found in a number of different domain architectures. The hypothesis is that the members of this superfamily have been shuffled among different proteins mainly by exon shuffling through intronic recombination [7,27]. Exon shuffling is a process of insertion of exon(s) coding for new domains in the ancestral gene introns. Domain shuffling by this mechanism is, in general, considered to be a powerful means that has shaped metazoan extracellular proteins [7]. Moreover, there have been two major changes in the rate of evolution of multidomain proteins: first in the move from unicellular to multicellular organization, and, secondly, within the vertebrate lineage [28], and Patthy [7] has proposed that these are related to the increase in the size and number of introns in the metazoan lineage. Namely, he suggests that the greater portion of introns in the animal genomes has made exon and consequentially domain shuffling easier. Evidence in the favour of this theory is also a finding that domains whose boundaries strongly correlate with exon boundaries exhibit significant expansion during animal proteome evolution, and have a great number of domain partners with which they occur [29].

Evidence of exon shuffling by which domains can be inserted into a protein, as well as the existence of alternative splicing by which exons and hence domains can be easily excluded from a protein, indicates that mechanisms that can add and delete domains from the middle of proteins are active in animals. However, there is a clear bias for changes to occur at the protein termini. The observed bias can partly be attributed to the higher frequency of other mechanisms which act at the protein termini. However, selective constraint imposed by the necessity for structural stability also favours changes at the termini and should not be ignored.

Of course, a protein's function and evolution is defined not only by its sequence, but also by its genomic position, expression pattern, and partners in its interaction network [30]. Although we know that evolution by natural selection plays with protein domains and has built a vast array of molecular machines, we still have a quite poor understanding of how and when domains are combined and reordered. We see that, in the coming years, with increasing access to population-specific data along with the huge diversity of complete animal genomes, we will gain a better understanding of how the various mutational mechanisms cause changes in protein domain architecture.

Acknowledgement

We thank Penny Coggill for critically reading the paper.

Funding

This work was supported by the Wellcome Trust [grant number WT 077044/Z/05/Z].

References

1 Holm, L. and Sander, C. (1994) Parser for protein folding units. Proteins **19**, 256–268

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

3 Babushok, D.V., Ostertag, E.M. and Kazazian, Jr, H.H. (2007) Current topics in genome evolution: molecular mechanisms of new gene formation. Cell. Mol. Life Sci. **64**, 542–554

4 Vogel, C. and Chothia, C. (2006) Protein family expansions and biological complexity. PLoS Comput. Biol. **2**, e48

5 Das, S. and Smith, T.F. (2000) Identifying nature's protein Lego set. Adv. Protein Chem. **54**, 159–183

6 Moore, A.D., Bjorklund, A.K., Ekman, D., Bornberg-Bauer, E. and Elofsson, A. (2008) Arrangements in the modular evolution of proteins. Trends Biochem. Sci. **33**, 444–451

7 Patthy, L. (1999) Genome evolution and the evolution of exon-shuffling: a review. Gene **238**, 103–114

8 Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature **402**, 86–90

9 Pasek, S., Risler, J.L. and Brezellec, P. (2006) Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics **22**, 1418–1423

10 Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein–protein interactions from genome sequences. Science **285**, 751–753

11 Basu, M.K., Carmel, L., Rogozin, I.B. and Koonin, E.V. (2008) Evolution of protein domain promiscuity in eukaryotes. Genome Res. **18**, 449–461

12 Meier, S., Jensen, P.R., David, C.N., Chapman, J., Holstein, T.W., Grzesiek, S. and Ozbek, S. (2007) Continuous molecular evolution of protein-domain structures by single amino acid changes. Curr. Biol. **17**, 173–178

13 Pils, B. and Schultz, J. (2004) Inactive enzyme-homologues find new function in regulatory processes. J. Mol. Biol. **340**, 399–404

14 Bartlett, G.J., Borkakoti, N. and Thornton, J.M. (2003) Catalysing new reactions during evolution: economy of residues and mechanism. J. Mol. Biol. **331**, 829–860

15 Wagner, A. (2002) Selection and gene duplication: a view from the genome. Genome Biol. **3**, reviews1012

16 Romo-González, T. and Vargas-Madrazo, E. (2006) Substitution patterns in alleles of immunoglobulin V genes in humans and mice. Mol. Immunol. **43**, 731–744

17 Bjorklund, A.K., Ekman, D., Light, S., Frey-Skott, J. and Elofsson, A. (2005) Domain rearrangements in protein evolution. J. Mol. Biol. **353**, 911–923

18 Weiner, 3rd, J., Beaussart, F. and Bornberg-Bauer, E. (2006) Domain deletions and substitutions in the modular protein evolution. FEBS J. **273**, 2037–2047

19 Li, H., Coghlan, A., Ruan, J., Coin, L.J., Heriche, J.K., Osmotherly, L., Li, R., Liu, T., Zhang, Z., Bolund, L. et al. (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. Nucleic Acids Res. **34**, D572–D580

20 Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000) The Pfam protein families database. Nucleic Acids Res. **28**, 263–266

21 Sammut, S.J., Finn, R.D. and Bateman, A. (2008) Pfam 10 years on: 10,000 families and still growing. Brief. Bioinform. **9**, 210–219

22 Fitch, W.A. (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**, 406–416

23 Zhang, J. (2003) Evolution by gene duplication: an update. Trends Ecol. Evol. **18**, 292–298

24 Harpaz, Y. and Chothia, C. (1994) Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. J. Mol. Biol. **238**, 528–539

25 Gamulin, V., Rinkevich, B., Schacke, H., Kruse, M., Muller, I.M. and and Muller, W.E. (1994) Cell adhesion receptors and nuclear receptors are highly conserved from the lowest metazoa (marine sponges) to vertebrates. Biol. Chem. Hoppe Seyler **375**, 583–588

26 Fraser, J.S., Yu, Z., Maxwell, K.L. and Davidson, A.R. (2006) Ig-like domains on bacteriophages: a tale of promiscuity and deceit. J. Mol. Biol. **359**, 496–507

27 van Rijk, A. and Bloemendal, H. (2003) Molecular mechanisms of exon shuffling: illegitimate recombination. Genetica **118**, 245–249

28 Ekman, D., Bjorklund, A.K. and Elofsson, A. (2007) Quantification of the elevated rate of domain rearrangements in metazoa. J. Mol. Biol. **372**, 1337–1348

29 Liu, M., Walch, H., Wu, S. and Grigoriev, A. (2005) Significant expansion of exon-bordering protein domains during animal proteome evolution. Nucleic Acids Res. **33**, 95–105

30 Pal, C., Papp, B. and Lercher, M.J. (2006) An integrated view of protein evolution. Nat. Rev. Genet. **7**, 337–348

Received 19 January 2009
doi:10.1042/BST0370751