CMLS, Cell. Mol. Life Sci. 62 (2005) 435–445
1420-682X/05/040435-11
DOI 10.1007/s00018-004-4416-1
© Birkhäuser Verlag, Basel, 2005

CMLS Cellular and Molecular Life Sciences

Review

The evolution of domain arrangements in proteins and interaction networks

E. Bornberg-Bauer[a], F. Beaussart[a], S. K. Kummerfeld[b], S. A. Teichmann[b] and J. Weiner III[a,]*

[a] Bioinformatics Division, School of Biological Sciences, University of Münster, Schlossplatz 4, 48143 Münster (Germany), e-mail: january@uni-muenster.de
[b] MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH (United Kingdom)

Received 20 September 2004; received after revision 23 October 2004; accepted 1 November 2004

Abstract. Proteins are composed of domains, which are conserved evolutionary units that often also correspond to functional units and can frequently be detected with reasonable reliability using computational methods. Most proteins consist of two or more domains, giving rise to a variety of combinations of domains. Another level of complexity arises because proteins themselves can form complexes with small molecules, nucleic acids and other proteins. The networks of both domain combinations and protein interactions can be conceptualised as graphs, and these graphs can be analysed conveniently by computational methods. In this review we summarise facts and hypotheses about the evolution of domains in multi-domain proteins and protein complexes, and the tools and data resources available to study them.

Keywords. Sequence analysis; domain evolution; regulatory networks; bioinformatics.

Introduction

Proteins and Domains

Proteins are the products of genes, and as such their evolution occurs by genetic mechanisms such as duplication and divergence (see recent review by Hurles [1]), recombination [2], insertions and deletions. Despite these processes that alter genes and protein sequences, close evolutionary relationships are apparent by comparison of DNA sequences. Since the 20-letter amino acid alphabet of proteins is more complex than the 4-letter alphabet of DNA, more distant evolutionary relationships are easier to detect by comparing protein sequences than DNA sequences. Even more divergent relationships can be observed at the level of protein three-dimensional structure (3D), because selection for functional characteristics occurs at this level, so protein structure is more conserved than sequence.

In the early days of the solution of protein structures, it became apparent that the same globular unit, or domain, can recur in different proteins with different partner domains. For instance, Michael Rossmann recognised the NADH-binding domain (named after him) in several dehydrogenases as early as 1974 [3]. As more protein structures were solved, classifications of domains and their evolutionary relationships were carried out (e.g. [4, 5]). While earlier hypotheses suggested that this recurrence is mainly to increase the functional versatility of proteins, recent insights have suggested that there may be fundamental biophysical causes underlying such rearrangements [6].

It is also important to recognise rearrangements of domains, because most sequence comparison methods break down when the linear order of domains is not maintained. Furthermore, the particular domain combination of a protein is crucial for its function and its participation in biological networks such as the protein interaction network of a cell [7]. Networks have emerged as a useful way of conceptualising the interplay of biomolecules, and

* Corresponding author.

can be formalised as graphs. Since standard methods exist to analyse such graphs, the deluge of 'omics' data has triggered the emergence of a completely new discipline, that of biological graph analysis [8–10].

In this review, we will summarise the current status of research and main insights on domain combinations in proteins. We will discuss the mechanisms and genome-wide characteristics of domain fusions, fissions, recombinations, insertions and losses. We will also briefly summarise bioinformatics tools and data resources to study these rearrangements. Recent insights into how networks evolve and, in particular, what role the rearrangement of domains plays in the emergence of protein interaction networks will be discussed in the last section.

Definitions of domains

A domain can be a whole small protein, or be part of medium-sized or larger proteins in combination with other domains. The most common and useful definition of domains is evolutionary. Domains that belong to the same family either have significant sequence similarity or similar 3D structures. An example of a domain database founded on sequence families is the Pfam database [11], while one that classifies 3D structures and thus includes more distantly related domains is the SCOP database [12]. SCOP domains usually have 100–250 residues, with the average length of a protein domain being around 175 amino acids [13]; Pfam domains are usually shorter, with about 145 amino acids.

Generally, proteins tend to be less conserved at the level of sequences than at the level of structure. Still, domains and conserved sequence motifs – such as PROSITE patterns or HMM models of Pfam – often coincide. However, the overlap of a structural domain and a conserved sequence motif depends on the chosen definition of both terms (see fig. 1). We will use the word 'domain' as a generic term to denote any conserved unit and assume that, from the context, it is always possible to infer the level of description.

Proteins can comprise a single domain or be made from several domains, resulting in a multi-domain protein. Structural assignments to gene sequences from complete genomes reveal that about two-thirds of prokaryotic proteins and 80% of eukaryote proteins are multi-domain proteins [14]. The increase in the complexity of the sequential order of domains in eukaryotic proteins, i.e. their domain architectures, has been termed domain accretion by Koonin and colleagues [15]. In multi-domain proteins in both prokaryotes and eukaryotes, most domains are formed by a continuous stretch of amino acids. There are exceptions, where one domain is inserted within another [16].

Based on their sequence and structural conservation, domains can be classified into families or superfamilies

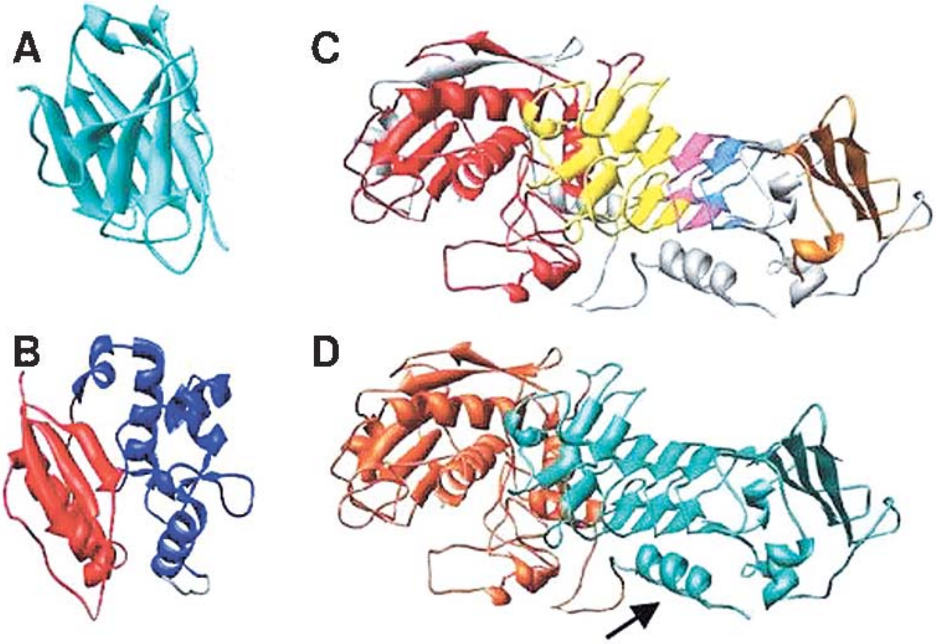

Figure 1. Examples of folds, domains and motifs. (*A*) Sandwich fold of an antibody (PDB 1a7n). (*B*) A two-domain protein (as annotated in the CATH database), human methyl-transferase *agt*; (PDB 1eh6) blue, an orthogonal bundle; red, a roll, (*C*) and (*D*) show a bacterial alkaline protease (1kap). (*C*), colour coded according to PFAM domain annotation (red, Peptidase_M10; yellow, Pfam-B_1569; pink and light blue, HemolysinCabind; orange, Pfam-B_3110), (*D*) According to protein fold annotation from the CATH database; turquoise, a discontinuous domain, a two-solenoid, arrow shows a fragment from a distant sequence region, but belonging to the two-solenoid fold; red, a three-layer ($\alpha\beta\alpha$) sandwich.

of related members. The SCOP database of structural domains contains around 1300 superfamilies at present [12], the CATH structural classification database has 3946 homologous families [17] and the Pfam database of sequence families contains about 7500 families [11]. Presumably, proteomes have evolved from a limited repertoire of domain families, and multi-domain proteins are assembled from combinations of domains from these families. For example, in the small-molecule pathways of *Escherichia coli*, around 90% of the roughly 600 enzymes are built from about 213 protein domain families [18]. Within this set of enzymes, as within individual genomes, there are domain families that consist of a single domain as well as large families that have many domain members. Domain families evolve by gene duplication and recombination, and this has resulted in a distribution of family sizes within each genome.

Family sizes in genomes are observed to follow power laws, with a few large families and many small families (e.g. [19]). The origin for this is a combination of neutral evolutionary events as well as selection for the functional properties of particular families. Moreover, the larger a genome in terms of the number of protein-coding genes, the higher the fraction of domains that are duplicated. 58% of all domains in *Mycoplasma* and 98% of all domains in human have been found to be duplicates [20–22].

Given this variety of large and small families within a genome, it should be noted that domains from different protein families can have the same structural arrangement or topology of secondary structure elements. The term ‘protein fold’ describes such similarity at the structural level without implying an evolutionary relationship between domains. So folds are a higher level of classification of domains that can encompass one or more superfamilies (see fig. 1C, D). This convergence of domains with different evolutionary origins and no sequence similarity to the same fold is due to the fact that the physical constraints favour certain arrangements or topologies of chains [23]. Folds in turn can be classified according to their secondary structure content such as all-alpha, all-beta etc.

The preponderance of multi-domain proteins in the three kingdoms of life underscores their role in the evolution of diverse molecular functions, and thus the understanding of their evolution becomes crucial. The types of associations, their classification and evolutionary background will be discussed in the following sections.

Conservation and variation of domain combinations

Just how versatile are domains in their combinations?

As early as the 1970s, it became clear that domains from the same family can combine with different neighbouring domains within a polypeptide chain. The three-dimensional structures of different dehydrogenases that all shared an NAD-binding Rossmann domain, but contained different catalytic domains, provided a clear example of this [3]. As more protein sequences and structures were determined experimentally, domains were defined and classified into families in a systematic way [4, 5, 24, 25].

With a large number of complete genome sequences available, we can gain an overview of the characteristics of domain combinations in multi-domain proteins through domain assignments to the proteins. The domain assignments are generated by identifying regions of homology between domains defined according to sequence and structural properties, and the proteins in the genomes. Homologies can be detected using powerful multiple sequence comparison methods such as hidden Markov models. Databases of hidden Markov models of domains and their protein families are available that also contain pre-computed domain architectures for proteins. Examples of such databases are Pfam [11], Superfamily [26], PSSM [27] and Gene3D [28].

The domain architectures of all multi-domain proteins for a given genome can be analysed in terms of pairs of domains that are adjacent to each other in the different proteins. This approach has revealed several overall features of domain combinations in multi-domain proteins [29]. First, between 10% and a quarter of all families occur as two or more tandem copies of the same type of domain which are thought to have evolved by internal gene duplication; these stretches of tandem domains are longer in eukaryotes than prokaryotes. Second, families that are more abundant and have more members within a genome also tend to have a greater variety of neighbouring domains. Therefore, the largest families in a genome also tend to be the most versatile in terms of their N- and C-terminal neighbours. For instance, the third-largest family in the human genome according to the Superfamily database (v. 1.65 [26]), the P-loop nucleotide triphosphate hydrolases, has 114 different types of adjacent domains. The seventh largest family, the protein kinase-like domains, has 82 different types of domain neighbours. There are a few such large families with many different types of neighbours. However, most families are small and have only one or two different combination partners.

The distribution of the number of different types of neighbours for each family follows a power law [30–32], as does the distribution of number of domain combinations for a given family. Even within a single family, this property of being scale-free is preserved. So although there are 114 different combinations of domain pairs with P-loop nucleotide triphosphate hydrolase domains in the human genome, about one-tenth of the proteins are combinations with a Translation protein domain. Similarly, for the Protein kinase-like domains, over one-fifth of the proteins within all the 82 different types of pairwise combinations are accounted for by SH2 domains. (Note that these numbers exclude proteins with tandem repeats of the two families.) Thus each family has a single or small group of neighbouring domains that dominate, and all other types of domain combinations occur in only a small number of proteins [Vogel et al., J. Mol. Biol., in press]. Furthermore, each domain pair occurs in only one N- to C-terminal order; about 5–10% of domain pairs are exceptions and occur in both orientations. The origin of this conservation of sequential order of domains will be discussed below.

It is evident that the answer to the question ‘Just how versatile are domains in their combinations?’ is that there is indeed a remarkable extent of domain shuffling in the protein repertoire. At the same time, duplication of domain combinations is very important, and new combinations of domains are more likely to arise by duplication of existing combinations than by recombinations. Thus there are far fewer different types of domain pairs than would be expected by random suffling of domains [33], and conservation of domain combinations occurs in proteins at several levels.

Conservation and variation of domain associations

As mentioned above, some domain combinations occur in several proteins within a genome. There can be two reasons for this: either there are several proteins with exactly the same domain architectures, or there are proteins with different domain architectures that have two or more sequential domains in common. Proteins with the same domain architecture, i.e. the same sequential order of domains from the N- to the C-terminus of the polypeptide chain, are likely to have evolved by gene duplication. This is supported by sequence and structural evidence [33, 34]. The same type of evidence suggests that when a subset of adjacent domains is shared between two proteins, this is due to evolutionary conservation of those domains.

This principle is illustrated by the P-loop nucleotide triphosphate hydrolase domains that are adjacent to Translation protein domains in 26 proteins in the human genome. These proteins, which are mostly translation factors, actually belong to eight different domain architectures and contain domains from four other families in addition to the P-loop and Translation protein domains. This example exhibits both conservation of a domain combination that has particular functional properties, as well as versatility in generating new domain architectures. Domain pairs or triplets that are conserved across domain architectures because they have functions that can be adapted to different domain contexts have been termed supra-domains by Vogel et al. [21].

The full extent of the conservation of domain combinations within the protein repertoire, both due to duplication of entire genes and due to conservation of parts of genes, is evident in the fact that less than 1% of all possible domain combinations are observed in multi-domain proteins with assigned domain architectures. So although there are over 9114 combinations of SCOP domains observed in 150 completely sequenced genomes, there is strong evolutionary conservation of domain combinations by duplication and modification of domains through sequence divergence. This also accounts for the conservation of N- to C-terminal order of domains: the same domain pair recurs by duplication and divergence as a single unit, rather than by independent recombination events that can result in inversion of the domain order.

So far, we have discussed domain combinations in terms

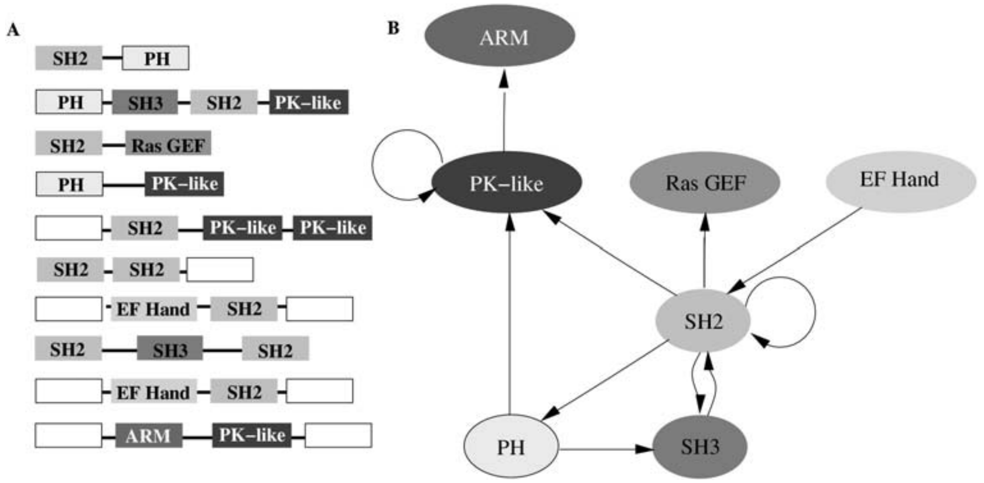

Figure 2. (*A*) A set of domain architectures with each coloured rectangle representing a domain. (*B*) The corresponding graph: ovals represent domain superfamilies (nodes in the graph), and edges indicate N-C terminal arrangement of domains within proteins. Extract is taken from *Homo sapiens* and shows a subset of proteins that are involved in signal transduction networks.

of linear strings of domains and adjacent domain pairs. A useful way of viewing the ensemble of adjacent domains in a genome or across several genomes is to represent each family as a node in a graph, and to connect families that are adjacent to each other, or that occur as part of the same protein. Formally, the domain graph G is given by G = (V, E). Each of the nodes or vertices V_i is a domain family, e.g. the Protein kinase or Rossmann domains. An edge $E_{i,j}$ denotes the adjacency of domains or the co-occurrence of domains within a protein that contains both the domains i and j, irrespective of their order. Depending on whether the edges represent adjacent domains or just co-occurrence of domains, two different graphs are obtained. Both types of graphs are scale-free networks, meaning that the number of different types of partner domains for each family is distributed according to a power law as shown for different domain definitions [31, 32], for structural domains [29] and some simplified model systems [32].

Ye and Godzik [35] developed a tool to visualise and compare domain co-occurrence graphs. They showed that these graphs contain clusters of closely connected domains. A graph approach has also been used by us to study the domain adjacency graph of eukaryote and prokaryote genomes [S. K. Kummerfeld and S. A. Teichmann, unpublished results]. In this graph, the edges between families are directed, depending on their N- to C-terminal order. If two domains of the same type are adjacent to each other, a node is connected to itself as a loop. For the most part, genome-wide domain graphs are sparsely connected: most nodes have few incoming or outgoing connections, and any given pair of domains has a single link in one direction or the other. However, some regions of these graphs are highly clustered, with connections back and forth between pairs of domains and their neighbours showing variation of the gene order. A recent study inspected the tightly interlinked regions of the graph in order to try to understand why they occur. It was found that the clusters often included functionally related domains with a particular bias towards certain types of function, including signal transduction and cell adhesion. For example, the cluster of domains shown in figure 2 are involved in signal transduction.

Mechanisms for generating new domain combinations

Fusion, fission and the Rosetta stone proteins

One way in which novel combinations can be formed is through duplication and rearrangement of domains. This process has been proposed to play a major role in the development of interaction interfaces [36, 37]. In essence, it is assumed that two proteins, typically single-domain proteins, are first fused into one gene. The physical interface between these two domains then evolves towards higher stability, e.g. by allowing for hydrophobic interactions that would not have evolved on the surfaces if the domains had remained separate. If, after many generations, the proteins are now split again, the interface remains as a recognition interface which may still mediate specific association of the two separated proteins. This process has also been termed the Rosetta stone method [38] since one protein, which still contains both domains, can occasionally be found in databases (see fig. 3). Rosetta stone proteins have been identified in various organisms.

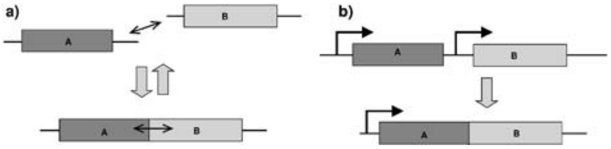

Figure 3. The Rosetta stone model. (*a*) A Rosetta stone protein as a fusion or fission of two proteins which interact with each other, and (*b*) as a fusion of genes, to allow a tighter co-regulation of both proteins.

Enright and Ouzounis [39] have followed up this idea and investigated 24 genomes for possible fusion events in an all-against-all comparison. The authors proposed 39,730 protein interactions for these genomes and compared these in silicio predictions with the results of yeast two-hybrid analysis. Only one pair of proteins from 957 putative interactions detected experimentally in yeast corresponded to a Rosetta stone protein.

However, the fact that a protein qualifies as a Rosetta stone protein does not necessarily need to be the proof of an interaction [40]. Two genes that are not directly interacting, but are co-regulated may undergo a fusion event that results in tighter co-regulation (see fig. 3b). This does not exclude an interaction either. Since many co-regulated genes are parts of multi-protein complexes, correlation between co-regulation and the presence of an interaction is to be expected. Such a correlation was indeed shown by Jansen et al. [41] for the interactions predicted by Enright and Ouzounis [39].

Another selective advantage may be genome reduction, a trend observed in several parasitic bacteria. Enright and Ouzounis [39] observed that in *Mycoplasma genitalium*, 15 Rosetta stone proteins correspond to distinct genes in the closely related *Mycoplasma pneumoniae*, which has a 17% larger genome; however, there are only 4 Rosetta stone proteins in *M. pneumoniae*. Thus, the selective force leading to genome reduction may also have led to the development of Rosetta stone proteins.

Differences between fusion and fission

A straightforward mechanism for domain rearrangements is the fusion of two genes to give a single, composite gene, or the reverse process, fission, where one gene is split into two or more separate components. Gene fusion is one way in which multi-domain proteins can arise during evolution. An example is the multi-functional fatty acid synthase complex, which has two polypeptide chains in yeast and mammals, but six chains in *E. coli*.

The majority of studies into gene fusion and fission have used sequence-based methods to identify groups of closely related proteins that have undergone fusion or fission, in order to predict functional relationships [36, 37, 39, 42]. Snel et al. [43] established the relative importance of fusion compared with fission by counting the number of proteins in split/component forms present in a set of prokaryotic proteins, concluding that fusion is more common than fission. A recent study has further quantified the relative rates of fussion and fission in multidomain proteins across the three kingdoms [S. K. Kummerfeld and S. A. Teichmann, Trends Genet., in press]. Proteins were considered in terms of their domain architectures and grouped into candidate fusion/fission sets. For example, proteins with the domain architecture CUB domain-EGF/laminin are grouped with single-domain CUB and EGF/laminin proteins from genomes that are missing the two-domain form. Assuming a species tree and placing the split/fused forms of proteins at the leaves of the tree, maximum parsimony is used to predict where in evolution fusion or fission events have occurred. This analysis showed that fusion is four times more common than fission, a consistent trend for Archaea, Bacteria and Eukaryotes.

Inversions and permutations of domains

As discussed above, domain achitecture is usually conserved, because duplication, which preserves domain order, is more common than recombination. In about 90% of the cases, any two domains always occur in the same order in protein sequences [29]. Rearrangements of domain order have an effect on the functionality of the protein if the interfaces between domains are important for the function of the protein, and if the rearrangement changes the geometry of the domain interface – for example, if the domains are packed in a different way. However, in vitro experiments show that several rearrangements of domain order have little or no effect on the functionality of the protein [44].

Circular or cyclic permutation (CP) is a type of nonlinear domain rearrangement such that an N-terminal fragment of one protein is similar to the C-terminus of an other protein, and vice versa (fig. 4a). Circular permutations have been described for various types of proteins. A well-described example is the DNA methyltransferases [45] and lectins [46]. Furthermore, CPs have been described for proteins as different as ABC transporters, ribosomal proteins, histones, homeobox proteins [47], oligopeptide binding proteins, chitinases, several dehydrogenases, bacterial antigens and many other proteins [48]. The existence of circular permutations lends support to the assumption that the function of protein domains depends most on the functionality of the domains and their physical (3D) arrangement, and not so much on the overall sequential domain arrangement of the protein. Also, CPs have been described to appear within many protein domains [49]. It is therefore of interest to study such rearrangements.

Two main mechanisms have been proposed for the origin of CPs (see fig. 4b, c): gene duplication followed by subsequent deletion events [45], and independent gene fusions [50]. In the first case (fig. 4b) the gene of concern first undergoes a duplication and an in-frame fusion. Such events are known to occur naturally. One or more duplications are then followed by subsequent insertions

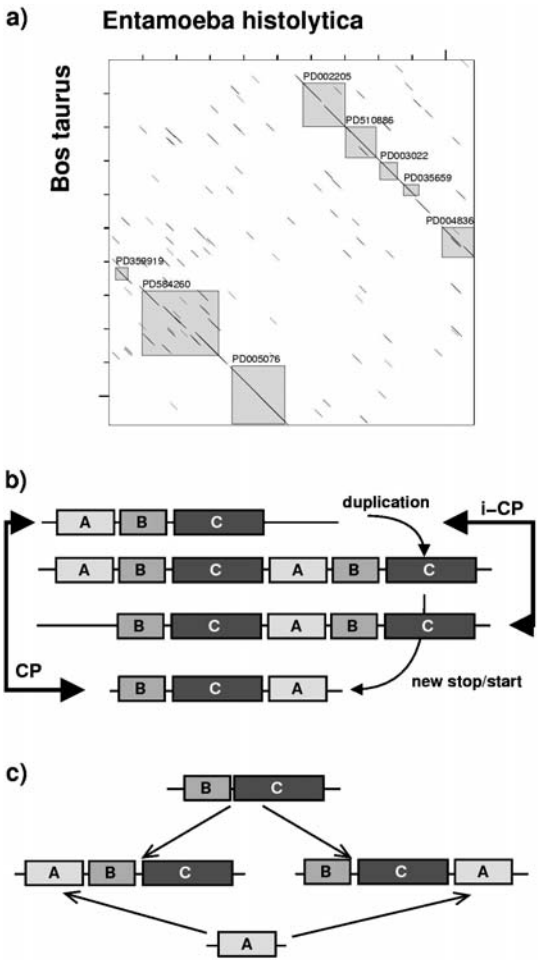

Figure 4. (*a*) Sequence and domain dot plot showing a circular permutation between two transhydrogenases. Horizontal, from the protozoan *Entamoeba histolytica*, vertical: bovine transhydrogenase. The squares correspond to domains that match between the two proteins. (*b–c*) mechanisms of CPs. (*b*) Circular permutations arising by gene duplication and (*c*) by independent gene fusion.

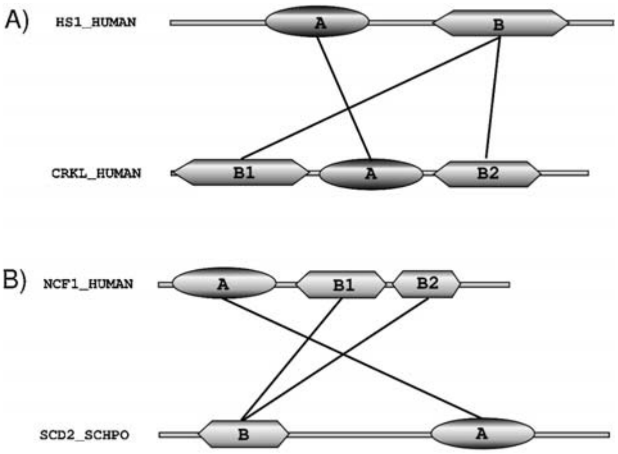

Figure 5. Examples of domain swaps containing domain duplications.(*A*) Domain B (hexagon) is duplicated in CRKL_HUMAN, becoming B1 and B2. If B1 is removed, no swap is detectable. Therefore, this is not a real domain swap. (*B*) domain B (hexagon) is duplicated, but there is still a swap.

of new start and stop codons. This mechanism implies the existence of intermediate states (intermediate circular permutations, iCPs, fig. 4b), with partly duplicated genes. These are difficult to discern from a real CP, but a recent algorithm working on strings of domains is supposed to solve this problem [48].

The second mechanism postulates the existence of proteins similar to the corresponding N- or C-terminal parts of the circularly permuted proteins. Independent fusion events of these genes may lead to formation of circularly permuted proteins (see fig. 4c). Such a case was found for the transhydrogenases from protozoans and higher Eukaryota.

A swap is similar to a CP in that a specific pair of domains can be found in reversed order but otherwise identical neighbourhood in two proteins (fig. 5). For instance, aXbYc and dY′eX′f contain a swap since X and X′ belong to one family (i.e. are the same type of domain) and Y and Y′ belong to another family. The intermittent regions (domains or linkers) a, b, c, d, e and f do not have to exist in all cases, and a pair of proteins can contain several swaps. Using this definition, several classes of swaps can be defined according to the different possible evolutionary scenarios [51]. Fliess et al. [51] designed algorithms to detect such swaps from SwissProt. As a main conclusion it appeared that swaps are rare, which lends further support to the assumption that maintenance of the relative order of domains is more important to retain a functional protein than the presence or absence of insertions or deletions.

This is also compliant with investigations on the fraction of reversals of pairs of adjacent domains. Reversals are reversed N- to C-terminal orientations of known domain combinations. Usually domain combinations appear in only one sequential order (AB or BA), but in rare cases (2%) both orientations can be found for a certain domain combination [34]. The majority of domain recombinations are independent combinations of existing domains (between 85 and 90% [33]); the fraction of true reversals is very low in domain combinations within genomes (1–4%) as well as across genomes (3–10% [30]).

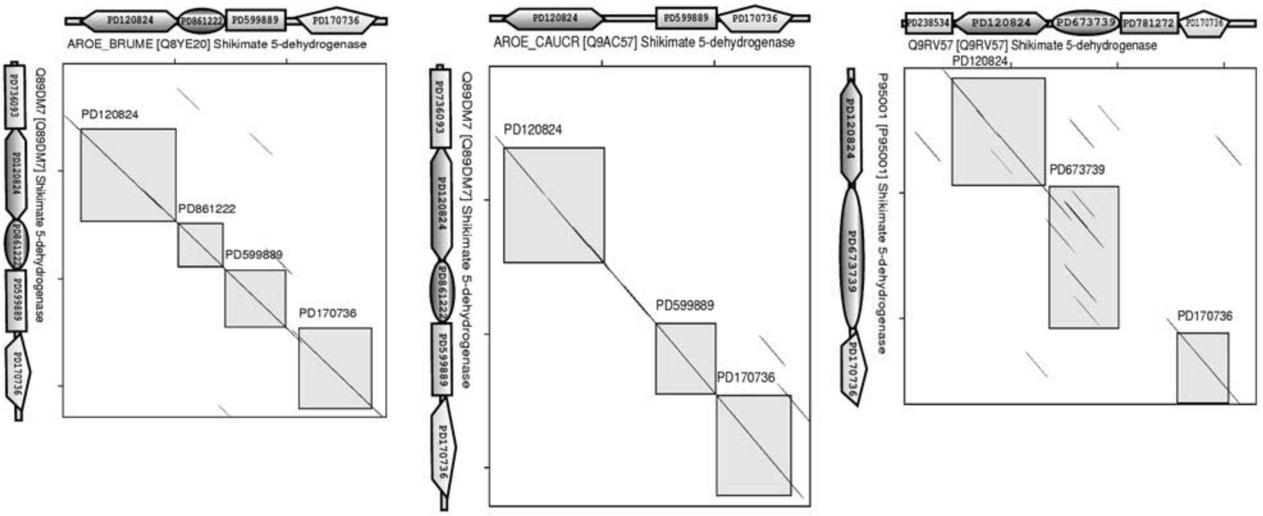

Figure 6. Examples of domain losses in shikimate 5-dehydrogenases. Grey squares denote matching domains; lines, amino acid sequence similarity. Left, (AROE_BRUME against Q89DM7): the first domain (PD736093, green rectangle) is physically lost in Q89DM7; Middle, (AROE_CAUCR against Q89DM7): the same as on the but the second domain in AROE_CAUCR (PD861222, ellipse) is also not annotated. However, it is not physically lost since there is still a significant trace in the dot-plot; Right, (9RV57 against P95001): two domains (PD238534, N-terminal rectangle and PD781272, rectangle) are physically lost in Q9RV57, although the latter may be wrongly annotated due to high similarities within domain PD673739. This is possibly the result of internal repeats or random matches.

Domain loss and domain insertions

Many pairs of proteins appear to have similar or almost identical domain order with just one exception, i.e. when a domain has been lost (or added) during evolution. Inference whether a loss or an addition have occurred can be done using a phylogenetic tree based on all common domains. Recent computations [F. Beaussart, unpublished data] suggested that domains towards the ends of a protein are lost more frequently than in the middle with an even stronger preference for the N-terminus. A closer inspection of these results suggests that a loss, as inferred by database annotation, does not always mean that the domain has been physically removed. Figure 6 suggests that some domains within a domain arrangement may have undergone stronger drift than others such that they are not picked up by an automated domain detection procedure such as ProDom. It is difficult to predict if such hidden domains are still functional or not. If the neighbouring domains still exist in the same order, it is most likely that the 'missing' domains may no longer fulfil their original function but have been retained because they still mediate some structural function which is essential for the function of the overall protein.

Insertions do not alway occur between two domains; a domain may also be placed within another domain. Investigations by Aroul-Selvam et al. [16] showed that such an insertion process is proportional to the length of the parent domain, i.e. the domain into which another one is inserted. Furthermore, considering the protein classes in the SCOP database, α/β proteins are more prone to have a domain inserted.

Versatility of domains in regulatory networks

One of the functions of domains is to mediate protein-protein interactions mentioned above. Domain versatility plays a great role in the creation of the next higher level of organisational units such as complexes and thus in the evolution of organisms in general. Such complexes are mainly mediated by protein-protein interactions and occur, for example, in transcription and signalling networks. It has been argued that the increasing complexity of multicellular organisms cannot just be the result of simply more genes. *Caenorhabitis elegans* has only around 1/3 less genes than humans, and the genome size of humans is only twice as large as the genome size of some Urochordates. To a large extent, this complexity is due to rewiring of networks and a higher complexity of regulatory networks. For example, the higher the organisation of an organism, the higher the proportion of transcription factors is [52]. Many transcription factors may be involved in the regulation of a single gene either alternatively or in concert, and each transcription factor can regulate many genes. In more complex organisms the average number of transcription factors which regulate, a

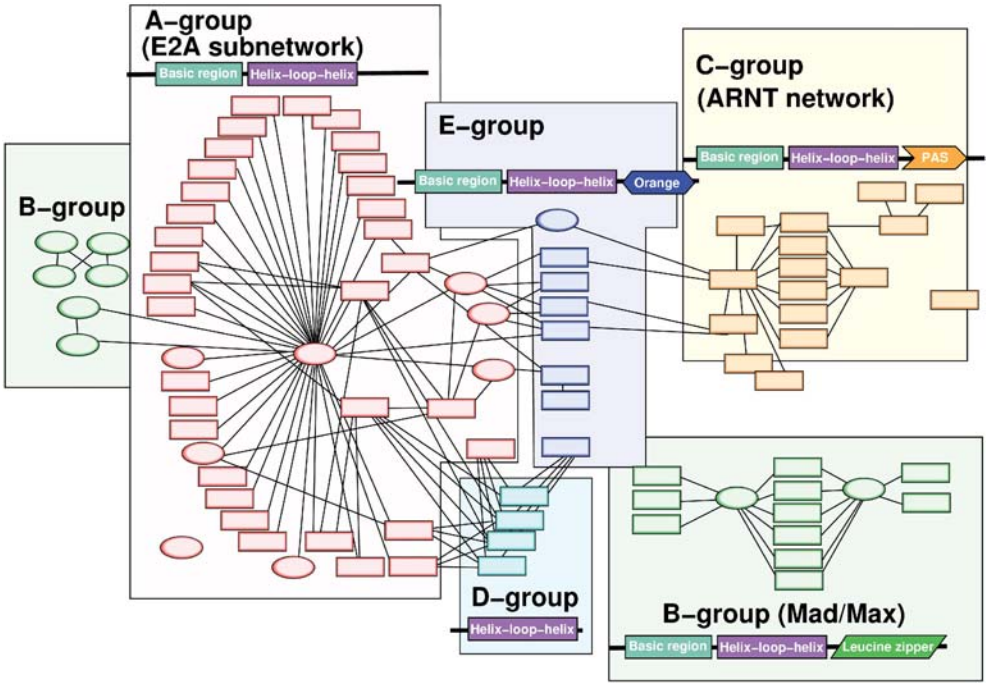

Figure 7. Interaction network of the transcription factors from the basic helix-loop-helix (bHLH) group. Ovals represent homodimerising proteins; boxes proteins, which can only heterodimerise; edges represent an existing interaction between two proteins. Different colours correspond to different bHLH families. The domain arrangement of the different bHLH families is shown underneath the corresponding family name (after Amoutzias et al. [7]).

Table 1. Tools and databases on domains available on the internet for analysis of domain composition and evolution in proteins.

Name	Purpose	Reference
Tools		
CADO	visualisation of protein domain organisation	Ye and Godzik [35]
DoMosaic	exploring the domain architecture in pairwise sequence comparisons	Gerrard and Bornberg-Bauer [62]
NIFAS	Java applet combining phylogenetic trees and domain arrangement visualisation	Storm and Sonnhammer [63]
RASPODOM	Identification of circular permutations	Weiner et al. [48]
TreeWiz	Inspecting and collapsing huge trees, displaying domain co-occurence	Rost and Bornberg-Bauer [64]
Databases		
BLOCKS	database of multiple alignments	Henikoff et al. [65]
Interpro	integrated resource for protein families, domains and functional sites	Apweiler et al. [66]
Pfam	highly annotated database of domains using HMM models	Bateman et al. [11]
PRINTS	collection of protein fingerprints (regular expressions)	Attwood [67]
PRODOM	database of domains generated automatically from SWISS-PROT and TrEMBL sequences	Corpet et al. [68]
PROSITE	database of protein sequence motifs (regular expressions)	Sigrist et al. [69]
SUPERFAMILY	database of domains using HMM models	Madera et al. [26]
SMART	resource for annotation of protein domains and analysis of domain architectures	Letunic et al. [70]
SCOP	structural classification of proteins	Murzin et al. [5]
CATH	hierarchical classification of protein domain structures	Orengo et al. [17]
FSSP	classification of 3D protein folds	Holm and Sander [71]
ProtoMap	classification of proteins by clustering into relational groups	Yona et al. [72]
ProClust	searching for homologue proteins by using transitivity	Pipenbacher et al. [73]
SYSTERS	large-scale protein clustering based on sequence similarity	Krause et al. [74]

Upper part: programs and web-based tools. Lower part: databases relevant for domain studies; sequence based, structure based and based on sequence clustering. The links to the web pages can be retrieved at http://www.uni-muenster.de/Biologie.Botanik/ebb/papers/review-domains/

gene is higher [53]. This rewiring of transcriptional circuits is in turn dependent on the modular rearrangement of the factor's constituents, such as proteins comprising domains interacting with either DNA or other proteins. It has been reported that DNA binding domains in many organisms descend from a relatively small, ancient core set which has been reused in varying combinations and frequencies depending on the lineage [54–56].

Similarly, protein-protein interaction networks, such as the ones in transcription factor complexes, change over evolutionary time and depend on combinatorial rearrangements of domains. Structural analysis showed that the number of domain interactions among proteins is limited in a fashion similar to domain combinations within proteins [57, 58].

For three families of eukaryotic transcription factor families, bHLH, NR and bZIP proteins, the interaction network has been investigated and combined with the analysis of their phylogenetic trees [7, 59, 60]. For the bHLH interaction network it was suggested that, beginning from initially homo-dimerising proteins, both homo-dimerising and hetero-dimerising proteins evolved primarily through series of single gene duplications, thus maintaining the homodimeric interaction while subsequently adding homologous heterodimeric interactions [7]. This gave rise to a hublike (star-shaped) network (fig. 7). More strikingly, networks are largely separated from each other since most interactions happen within one family. This family-specific interaction pattern correlates well with family-specific alternate arrangements with additional dimerisation domains. Since it has been repeatedly shown that such arrangements have subtle but important structural influences on dimerisation specificities, the modular domain rearrangement has, at least in this case, direct implications for the emergence of new networks. Similar conclusions hold for the family of nuclear receptor (NR) and bZIP proteins [59].

In summary, networks are a good example of how rearrangements of protein domains facilitate the emergence of higher-level cellular structures. Since the evolution of these networks can be investigated using traditional methods such as phylogenies, and the major driving forces are gene duplication, genetic drift and domain rearrangements, these insights have begun 'taking the mystery out of biological networks' [61].

Discussion

Masses of new genomic, proteomic etc. data arrive almost daily. While on the one hand the puzzle of domain versatility seems to become clearer, new insights about the actual role of domain versatility, such as for the emergence of networks, are just beginning to arise. We have summarised most of the recent insights. In essence, the mechanisms for domain rearrangements are manifold, and proteins appear to be very tolerant, in some evolutionary sense even dependent on more variation being generated in terms of modular rearrangements as long as the relative sequential order is essentially maintained.

Irregular arrangements can be cumbersome for the analysis of sequences using databases. However, the arrival of new bioinformatics tools, some of which we have summarised (table 1), may help to cope with such irregular cases. Consequently, new insights and exciting times exploring proteins and networks are just around the next corner. This should pave the way to a better understanding of complex biological systems and their evolution.

1 Hurles M. (2004) Gene duplication: the genomic trade in spare parts. PLoS Biol **2:** E206

2 Patthy L. (1999) Protein Evolution, Blackwell Science, Oxford

3 Rossmann M., Moras D. and Olsen K. (1974) Chemical and biological evolution of nucleotide-binding protein. Nature **250:** 194–199

4 Janin J. and Chothia C. (1985) Domains in proteins: definitions, location and structural principles. Methods Enzymol. **115:** 420–430

5 Murzin A., Brenner S., Hubbard T. and Chothia C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247:** 536–540

6 Cui Y., Wong W. H., Bornberg-Bauer E. and Chan H.S. (2002) Recombinatoric exploration of novel folded structures: a heteropolymer-based model of protein evolutionary landscapes. Proc. Natl. Acad. Sci. USA **99:** 809–814

7 Amoutzias G., Robertson D., Oliver S. and Bornberg-Bauer E. (2004a) Convergent evolution of gene networks by single-gene duplications in higher eukaryotes. EMBO Rep. **5:** 274–279

8 Barabasi A. and Oltvai Z. (2004) Network biology: understanding the cell's functional organization. Nat. Rev. Genet. **5:** 101–113

9 Spirin V. and Mirny L. (2003) Protein complexes and functional modules in molecular networks. Proc. Natl. Acad. Sci. USA **100:** 12123–12128

10 Bork P., Jensen L., von Mering C., Ramani A., Lee I. and Marcotte E. (2004) Protein interaction networks from yeast to human. Curr. Opin. Struct. Biol. **14:** 292–299

11 Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S. et al. (2002) The Pfam protein families database. Nucleic Acids Res. **30:** 276–280

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

13 Gerstein M. (1997) A structural census of genomes: comparing bacterial, eukaryotic and archaeal genomes in terms of protein structure. J. Mol. Biol. **274:** 562–576

14 Liu J. and Rost B. (2004) CHOP proteins into structural domain-like fragments. Proteins **55:** 678–688

15 Koonin E. (2000) How many genes can make a cell: the minimal-gene-set concept. Annu. Rev. Genomics Hum. Genet. **1:** 99–116

16 Aroul-Selvam R., Hubbard T. and Sasidharan R. (2004) Domain insertions in protein structures. J. Mol. Biol. **338:** 633–641

17 Orengo C., Michie A., Jones S., Jones D., Swindells M. and Thornton J. (1997) CATH-a hierarchic classification of protein domain structures. Structure **5:** 1093–1108

18 Teichmann S., Rison S., Thornton J., Riley M., Gough J. and Chothia C. (2001) Small-molecule metabolism: an enzyme mosaic. Trends Biotechnol **19:** 482–486

19 Qian J., Luscombe N. and Gerstein M. (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. J. Mol. Biol. **313:** 673–681

20 Teichmann S., Park J. and Chothia C. (1998) Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. Proc. Natl. Acad. Sci. USA **95:** 14658–14663

21 Vogel C., Bashton M., Kerrison N., Chothia C. and Teichmann S. (2004) Structure, function and evolution of multidomain proteins. Curr. Opin. Struct. Biol. **14:** 208–216

22 Muller A., MacCallum R. and Sternberg M. (2002) Structural characterization of the human proteome. Genome Res. **12:** 1625–1641

23 Chothia C. and Gerstein M. (1997) Protein evolution. How far can sequences diverge? Nature **385:** 579, 581

24 Patthy L. (1985) Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. Cell **41:** 657–663

25 Bork P. (1991) Shuffled domains in extracellular proteins. FEBS Lett. **286:** 47–54

26 Madera M., Vogel C., Kummerfeld S., Chothia C. and Gough J. (2004) The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res. **32:** D235-239

27 Kelley L., MacCallum R. and Sternberg M. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. J. Mol. Biol. **299:** 499–520

28 Buchan D., Rison S., Bray J., Lee D., Pearl F., Thornton J. et al. (2003) Gene3D: structural assignments for the biologist and bioinformaticist alike. Nucleic Acids Res. **31:** 469–473

29 Apic G., Gough J. and Teichmann S. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J. Mol. Biol. **310:** 311–325

30 Apic G., Gough J. and Teichmann S. (2001) An insight into domain combinations. Bioinformatics **17 Suppl. 1:** S83–89

31 Wuchty S. (2001) Scale-free behaviour in protein domain networks. Mol. Biol. Evol. **18:** 1694–1702

32 Bornberg-Bauer E. (2002) Randomness, structural uniqueness, modularity and neutral evolution in sequence space of model proteins. Z. Phys. Chem. **216:** 139–154

33 Apic G., Huber W. and Teichmann S. (2003) Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. J. Struct. Funct. Genomics **4:** 67–78

34 Bashton M. and Chothia C. (2002) The geometry of domain combination in proteins. J. Mol. Biol. **315:** 927–939

35 Ye Y. and Godzik A. (2004) Comparative analysis of protein domain organization. Genome Res. **14:** 343–353

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

37 Marcotte E., Pellegrini M., Yeates T. and Eisenberg D. (1999a) A census of protein repeats. J. Mol. Biol. **293:** 151–160

38 Marcotte E., Pellegrini M., Ng H., Rice D., Yeates T. and Eisenberg D. (1999b) Detecting protein function and protein-protein interactions from genome sequences. Science **285:** 751–753

39 Enright A. and Ouzounis C. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. Genome Biol **2:** RESEARCH0034

40 Veitia R. (2002) Rosetta Stone proteins: 'chance and necessity'? Genome Biol. **3:** INTERACTIONS1001

41 Jansen R., Greenbaum D. and Gerstein M. (2002) Relating whole-genome expression data with protein-protein interactions. Genome Res. **12:** 37–46

42 Yanai I., Derti A. and DeLisi C. (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. Proc. Natl. Acad. Sci. USA **98:** 7940–7945

43 Snel B., Bork P. and Huynen M. (2000) Genome evolution. Gene fusion versus gene fission. Trends Genet. **16:** 9–11

44 Lindqvist Y. and Schneider G. (1997) Circular permutations of natural protein sequences: structural evidence. Curr. Opin. Struct. Biol. **7:** 422–427

45 Jeltsch A. (1999) Circular permutations in the molecular evolution of DNA methyltransferases. J. Mol. Evol. **49:** 161–164

46 Young N., Williams R., Roy C. and Yaguchi M. (1982) Structural comparison of the lectin from sainfoin (*Onobrychis viciifolia*) with concanavalin A and other D-mannose specific lectins. Can. J. Biochem. **60:** 933–941

47 Uliel S., Fliess A. and Unger R. (2001) Naturally occurring circular permutations in proteins. Protein Eng. **14:** 533–542

48 Weiner III J., Thomas G. and Bornberg-Bauer E. (2004) Rapid motif-based prediction of circular permutations in multidomain proteins. Bioinformatics, in press

49 Jung J. and Lee B. (2001) Circularly permuted proteins in the protein structure database. Protein Sci. **10:**1881–1886

50 Bujnicki J. (2002) Sequence permutations in the molecular evolution of DNA methyltransferases. BMC Evol. Biol. **2:** 3

51 Fliess A., Motro B. and Unger R. (2002) Swaps in protein sequences. Proteins **48:** 377–387

52 Levine M. and Tjian R. (2003) Transcription regulation and animal diversity. Nature **424:** 147–151

53 van Nimwegen E. (2003) Scaling laws in the functional content of genomes. Trends Genet. **19:** 479–484

54 Madan Babu M. and Teichmann S. (2003) Evolution of transcription factors and the gene regulatory network in *Escherichia coli.* Nucleic Acids Res. **31:** 1234–1244

55 Perez-Rueda E. and Collado-Vides J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. Nucleic Acids Res. **28:** 1838–1847

56 Aravind L. and Koonin E. (1999) DNA-binding proteins and evolution of transcription regulation in the archaea. Nucleic Acids Res. **27:** 4658–4670

57 Park J., Lappe M. and Teichmann S. (2001) Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. J. Mol. Biol. **307:** 929–938

58 Russell R., Alber F., Aloy P., Davis F., Korkin D., Pichaud M. et al. (2004) A structural perspective on protein-protein interactions. Curr. Opin. Struct. Biol. **14:** 313–324

59 Amoutzias G., Robertson D. and Bornberg-Bauer E. (2004b) The evolution of protein interaction networks in regulatory proteins. Comp. Funct. Genom. **5:** 79–84

60 Amoutzias G.D., Weiner J. III and Bornberg-Bauer E. (2004c) Network phylogeny of protein intreactions in eukaryotic transcription factors, submitted

61 Aloy P. and Russell R. (2004) Taking the mystery out of biological networks. EMBO Rep. **5:** 349–350

62 Gerrard D. and Bornberg-Bauer E. (2003) doMosaic: analysis of the mosaic-like domain architecture in proteins. Informatica **27:** 15–20

63 Storm C. and Sonnhammer E. (2001) NIFAS: visual analysis of domain evolution in proteins. Bioinformatics **17:** 343–348

64 Rost U. and Bornberg-Bauer E. (2002) TreeWiz: interactive exploration of huge trees. Bioinformatics **18:** 109–114

65 Henikoff J., Greene E., Pietrokovski S. and Henikoff S. (2000) Increased coverage of protein families with the blocks database servers. Nucleic Acids Res. **28:** 228–230

66 Apweiler R., Attwood T., Bairoch A., Bateman A., Birney E., Biswas M. et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. **29:** 37–40

67 Attwood T. (2002) The PRINTS database: a resource for identification of protein families. Brief Bioinform. **3:** 252–263

68 Corpet F., Servant F., Gouzy J. and Kahn D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res. **28:** 267–269

69 Sigrist C., Cerutti L., Hulo N., Gattiker A., Falquet L., Pagni M. et al. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. **3:** 265–274

70 Letunic I., Copley R., Schmidt S., Ciccarelli F., Doerks T., Schultz J. et al. (2004) SMART 4.0: towards genomic data integration. Nucleic Acids Res. **32:** D142–144

71 Holm L. and Sander C. (1996) The FSSP database: fold classification based on structure-structure alignment of proteins. Nucleic Acids Res. **24:** 206–209

72 Yona G., Linial N. and Linial M. (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Nucleic Acids Res. **28:** 49–55

73 Pipenbacher P., Schliep A., Schneckener S., Schonhuth A., Schomburg D. and Schrader R. (2002) ProClust: improved clustering of protein sequences with an extended graph-based approach. Bioinformatics **18 Suppl. 2:** S182–191

74 Krause A., Nicodeme P., Bornberg-Bauer E., Rehmsmeier M. and Vingron M. (1999) WWW access to the SYSTERS protein sequence cluster set. Bioinformatics **15:** 262–263

To access this journal online:
http://www.birkhauser.ch