doi:10.1016/j.jmb.2005.08.067 J. Mol. Biol. (2005) 353, 911–923

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Domain Rearrangements in Protein Evolution

Åsa K. Björklund†, Diana Ekman†, Sara Light, Johannes Frey-Skött and Arne Elofsson*

Stockholm Bioinformatics Center, Stockholm University SE-10691 Stockholm, Sweden

*Corresponding author

Most eukaryotic proteins are multi-domain proteins that are created from fusions of genes, deletions and internal repetitions. An investigation of such evolutionary events requires a method to find the domain architecture from which each protein originates. Therefore, we defined a novel measure, domain distance, which is calculated as the number of domains that differ between two domain architectures. Using this measure the evolutionary events that distinguish a protein from its closest ancestor have been studied and it was found that indels are more common than internal repetition and that the exchange of a domain is rare. Indels and repetitions are common at both the N and C-terminals while they are rare between domains. The evolution of the majority of multi-domain proteins can be explained by the stepwise insertions of single domains, with the exception of repeats that sometimes are duplicated several domains in tandem. We show that domain distances agree with sequence similarity and semantic similarity based on gene ontology annotations. In addition, we demonstrate the use of the domain distance measure to build evolutionary trees. Finally, the evolution of multi-domain proteins is exemplified by a closer study of the evolution of two protein families, non-receptor tyrosine kinases and RhoGEFs.

Keywords: protein evolution; multi-domain proteins; proteome; GOGraph; Pfam

Introduction

Proteins are composed of domains, recurrent protein fragments with distinct structure, function and/or evolutionary history. Protein domains may occur alone, as single-domain proteins, but many are found in combination with other domains in larger polypeptide chains. These multi-domain architectures are more frequent in eukaryotes than prokaryotes.[1–4] During evolution, proteins with new functions or specificities have been invented through domain fusion and recombination as well as differentiation of existing domains. Domain fusion is a mechanism that allows the limited number of functional modules to be reused instead of reinvented. The occurrence of domain families as well as the number of partner families follow a power-law distribution with a few very abundant and/or versatile domains.[1,5] However, the evolution of domain combinations is not purely stochastic, but depends upon selection of certain functions.[6] Often two or three domains in tandem have been reused in combination with other domains. These supra-domains may have been selected because the function is dependent on the interface between them or because they are both necessary for proper function.[7] It has also been seen that some exon-bordering domains have unexpectedly many combination partners in animals.[8]

The addition of a domain to a protein is likely to alter its function, for example, it has been estimated that single-domain proteins from the same domain family have a 67% chance of having similar functions, whereas the corresponding number for two-domain proteins with just one of the domains in common is 35%.[9] Jensen proposed that ancient enzymes with broad substrate specificities have evolved into more specific enzymes through gene duplication.[10] Enzymes often retain their biochemical function while gaining new substrate specificities or regulation mechanisms by the addition of a domain. As a matter of fact, enzymatic function is conserved down to 30% sequence identity for most

† A.K.B. and D.E. contributed equally to this work.

Abbreviations used: Pfam, the protein family database; GO, gene ontology; DD, domain distance; DA, domain architecture.

E-mail address of the corresponding author: arne@sbc.su.se

single-domain enzymes and addition of a second domain rarely affects function.[11]

Sequence alignment based methods, such as ClustalW,[12] are often used to determine the evolutionary or functional relationship between proteins. However, multi-domain proteins may cause problems when creating multiple alignments. The sequences may align poorly for distantly related proteins even if they share the same domain architecture. A tool for finding related proteins based on domain architecture is CDART at NCBI[13] and another useful tool is NIFAS,[14] which is a domain evolution visualizer that builds trees based on the sequence alignments.

Understanding the underlying mechanisms of protein evolution through domain rearrangements and sequence differentiation is crucial for understanding the development of new functionalities. We have defined a new measure "domain distance", where each domain addition/deletion between two domain architectures is counted. We explore how domain distances correlate with sequence similarity and functional similarity. Using domain distances we have quantified the frequency of different events such as domain indels, repetitions and exchanges. These results were compared with frequencies obtained using a sequence based method. In addition, we demonstrate the possibility to use trees based on domain distance for exploring protein evolution. Finally, two protein families, the non-receptor tyrosine kinases and the RhoGEFs, serve as examples of domain rearrangements in protein evolution.

Domain Distance

It is well known that multi-domain proteins are created from fusions of whole or parts of genes and from internal duplications. In an attempt to quantify these events we have defined a novel measure of similarity between domain architectures (DAs), called domain distance (DD). Domain distance is calculated as the number of unmatched domains in an alignment of two architectures and is related to the number of evolutionary events required to evolve from one protein to another (see Figure 1). Naturally, the same DA may have been created more than once, but in most instances they have evolved from the same ancestor.[15] In addition, one event may consist of the insertion or deletion of several domains at a time; however, our results given below indicate that such events are not frequent. In order to avoid prior bias our model treats internal duplications (repetitions) and fusions equally, although repeats may be exposed to a different evolutionary pressure compared to fusions.[2,16]

Traditionally, phylogenetic studies use sequence similarity to identify homologues. However, distantly related sequences are difficult to detect with sequence-based methods, whereas domains can be detected more sensitively. Hence, domain distance may be an alternative to sequence similarity measures. Further, domain distances are not dependent on the length or conservation of domains, as sequence comparisons may be, and all domain families are treated equally, hence domain distances can group distant homologues even if they do not contain the "main" domain. The DD method is not, however, applicable to proteins

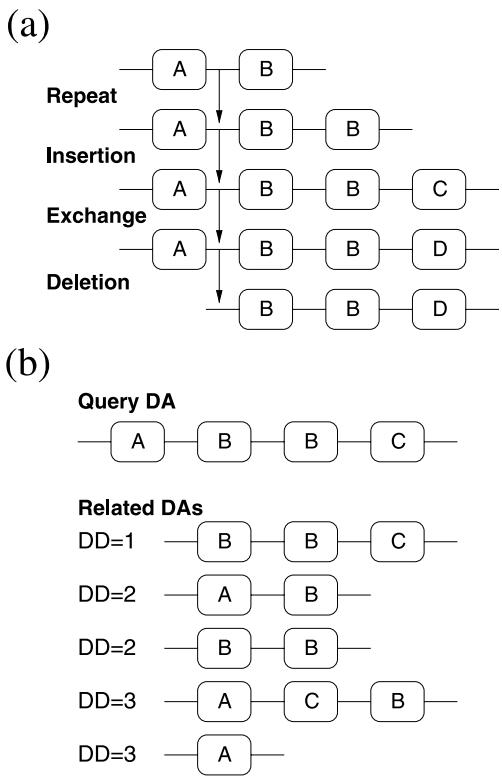

Figure 1. Evolutionary events and domain distance. (a) The evolutionary events that have created new domain architectures. Repeats are additions of new domain(s) from the same family as one of the adjacent domains, whereas an insertion is an addition of new domain(s) from a family/families other than the adjacent domains. Domains may also be deleted or exchanged. With regards to exchange of domains, where one domain has been substituted for another, we cannot tell if the domain has evolved to belong to a new domain family, if first one domain is deleted and then another is inserted or if, by some mechanism, they have been switched. In addition, examples of domains that have been inserted within other domains also exist[39] but as such examples are rare[2] we have chosen to ignore them. (b) Domain distance. The domain distance (DD) between two proteins is defined as the number of unmatched domains in an alignment between two proteins. For example, between the query and the domain architecture A-C-B, four domains are involved in the alignment (A-A and B-B), whereas three domains (B, C, C) are unmatched, hence DD = 3. The domain distance between two proteins with no domains in common is infinite. The DD between the query and all other domain architectures is calculated and the DA to which it has the shortest DD is defined as the closest neighbor, in this example BBC. Hence, an N-terminal insertion of one domain is counted.

with no domain families in common, since they have "infinite" distances. Neither does it separate proteins with identical domain architectures and hence is not suitable for studies of closely related proteins. However, for studying the evolution of new domain architectures, DD may complement sequence-based methods, as will be demonstrated below.

A general problem when studying domain architectures is that all domains cannot be found, either if they have evolved too far to be detected with current methods, or if they are not represented in the domain databases. Here, we mainly used Pfam-A domains instead of structurally defined SCOP domains, since they have better coverage in the genomes, especially for membrane proteins.[2] For comparison, though, we have also used SCOP. A disadvantage with Pfam is that different domain families may be of common origin, consequently, if a domain has diverged to another family, homologues will not be detected or the domain distance method will detect the event as a domain exchange. Many related domain families have been grouped together in the Pfam Clans†. The domains were reassigned using the clan classification; however, the number of multi-domain architectures was only slightly altered, since 86% of the clans were found in single-domain proteins, consequently the effect on the statistics of different events was also marginal.

As domain assignments are not perfect, large regions of the proteins lack domain assignments. To increase the coverage, we extended the Pfam-A domain assignments with Pfam-B, which yielded a larger number of DAs and events; however, the fraction of exchanges increased tenfold (data not shown). As was demonstrated previously,[2] many Pfam-B families are related to each other and to Pfam-A families and hence we are likely to overestimate the number of exchanges. Due to the insufficient clustering of Pfam-B families, the evolution of these regions has not been studied here. Naturally, we are not certain whether unassigned regions contain orphan domains or contain domains that have not been detected. Therefore, proteins with long unassigned regions were omitted from our calculations. This reduction obviously causes loss of domain architectures but also reduces the uncertainty in the dataset.

Domain distance *versus* sequence- and functional similarity

To determine the evolutionary distance between two multi-domain proteins we have employed the DD measure, instead of traditional sequence similarity. Here, we show how this measure relates to sequence similarity for a set of SWISS-PROT proteins with Pfam-A domain assignments (see Figure 2(a)). For short domain distances there is a correlation between sequence similarity and domain distances. However, at longer distances, which mainly derive from large multi-domain proteins with only one domain in common, the sequence similarity is approximately equal irrespective of DD, but always higher than for unrelated proteins. To investigate if the individual domains have diverged with increasing domain distance, the domain regions were aligned separately. This gave similar results as aligning the whole protein, however, at long domain distances the bit score decreased more for the domain alignments. It was clear that for some domain families, such as the leucine-rich repeats Krab and SH2, the conservation decreased with increasing DD. However, for other families, such as C2H2 zinc fingers and protein kinase domains, an increase in domain distance did not correspond to decreasing sequence similarity and only at DD zero was the sequence conservation higher than at longer distances.

As many domains are functional units, DD was also compared with semantic similarity based on GO annotations. Semantic similarity has been demonstrated to correlate well with sequence similarity,[17] and as shown in Figure 2(b), there is a good agreement between semantic similarity and DD. As can be expected, proteins with at least one domain in common have higher semantic similarity than unrelated proteins. Domain distance seems to be especially well correlated with molecular function, while the correlation is weaker for biological process and cellular component. It is notable that all functional similarities are higher at distance 1 than at distance zero. This is due to the large family of seven transmembrane receptors that account for two-thirds of all pairs of single-domain proteins, which all have the same domain family but varying functional annotations.

Which are the most common domain rearrangements?

The evolution of multi-domain proteins can be studied by the use of phylogenetic trees built from sequence alignments. However, to obtain a picture of the individual events behind the creation of novel multi-domain proteins we have used domain distances to find the closest neighbor for each domain architecture.

Here, evolutionary events are categorized into three different classes, repetitions, indels (insertions/deletions) and domain exchanges, see Figure 1(a). We have not separated insertions from deletions, as this is not possible using domain architectures only. Therefore, we assume that each multi-domain architecture has arisen from another architecture that is shorter or of the same length, since it has been shown that more proteins are created from fusion (insertion) than fission (deletion)[18] and that fusions are four times more common.[19]

Frequencies of the events required for creating a domain architecture from its closest neighbor were calculated on two datasets, SWISS-PROT and a set

† ftp://ftp.sanger.ac.uk/pub/databases/Pfam/.

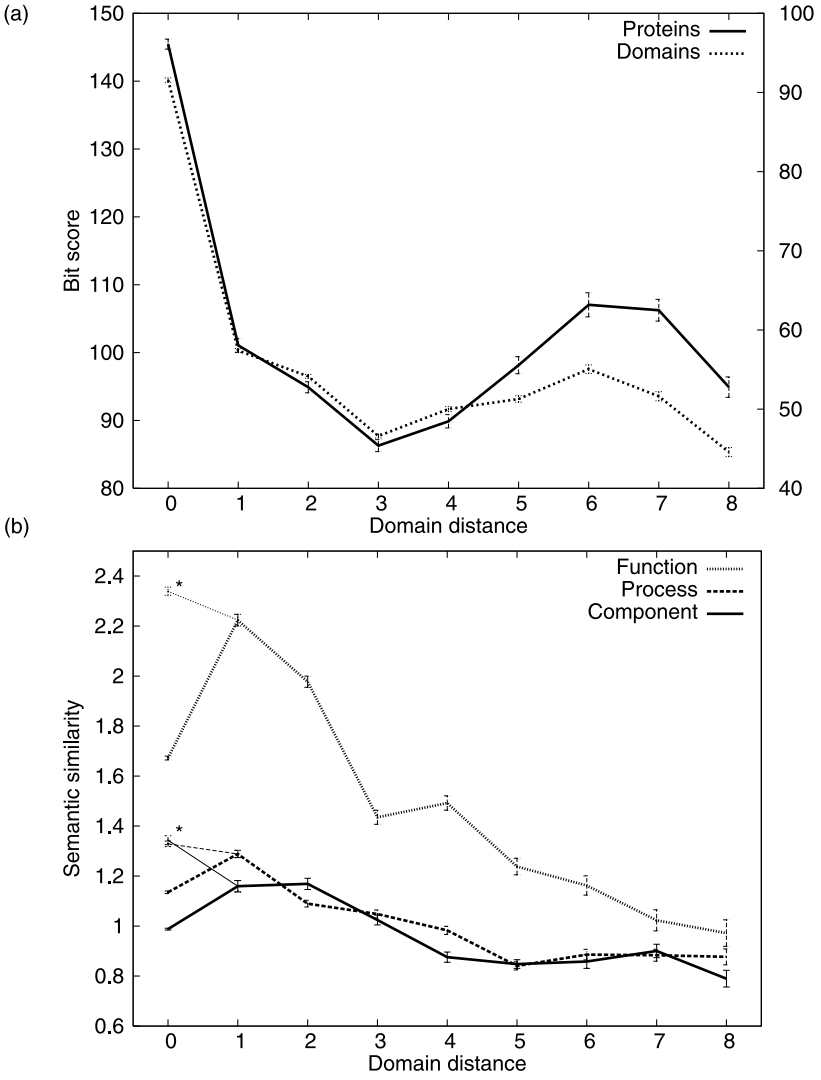

Figure 2. (a) Sequence bit scores at different domain distances calculated for all pairs in a non-redundant set from SWISSPROT with only one representative for each domain-combination and only containing proteins that have GO-annotations (TAS). Sequence bit scores were binned together at intervals of 10. For proteins with no domains in common, i.e. with infinite domain distances the average bit score was 17.1 ± 0.1. (b) Semantic similarity (GOGraph) for protein pairs with different domain distances. Only non-redundant protein pairs with GO-annotation (TAS) were compared and average semantic similarity for Molecular Function, Biological Process and Cellular Component were calculated at each domain distance. For proteins with infinite domain distance, the average semantic similarity was 0.57 ± 0.02 for Function, 0.69 ± 0.02 for Process and 0.52 ± 0.02 for Component. (*) The low semantic similarities at DD=0 can be explained by the large group of seven transmembrane receptors with just one 7tm domain (PF00001) but differing functions, with the exclusion of all such proteins, the higher semantic similarities indicated with a star are obtained.

consisting of seven eukaryotic proteomes. As can be seen in Table 1, indels and repetitions are frequent, while exchanges of domains are very rare. When determining the N or C-terminal position of an event, only architectures with three or more domains were considered, since two-domain proteins are often created from two single-domain proteins and, as a result, the position of the insertion cannot be determined. This reduced the datasets to 958 Pfam DAs and 758 SCOP DAs in the eukaryotic

Table 1. Summary of evolutionary events for: eukaryotes using domain distance (with Pfam and SCOP domains), SWISS-PROT Pfam using domain distance and SWISS-PROT Pfam using sequence alignments (SWISS-PROT seq)

Dataset	Event	Number	Fraction
Eukaryotes Pfam	Ind	1147	0.67
	Rep	746	0.43
	Exc	42	0.02
Eukaryotes SCOP	Ind	1535	0.76
	Rep	597	0.29
	Exc	14	0.01
SWISS-PROT Pfam	Ind	1132	0.77
	Rep	530	0.36
	Exc	76	0.05
SWISS-PROT seq Pfam	Ind	1159	0.74
	Rep	683	0.44
	Exc	244	0.16

The number of occurrences and the fraction of domain architectures for each of the three events (indels, repetitions and exchanges) are presented. Since more than one event may have occurred between two architectures, the sum of all fractions may exceed 1.

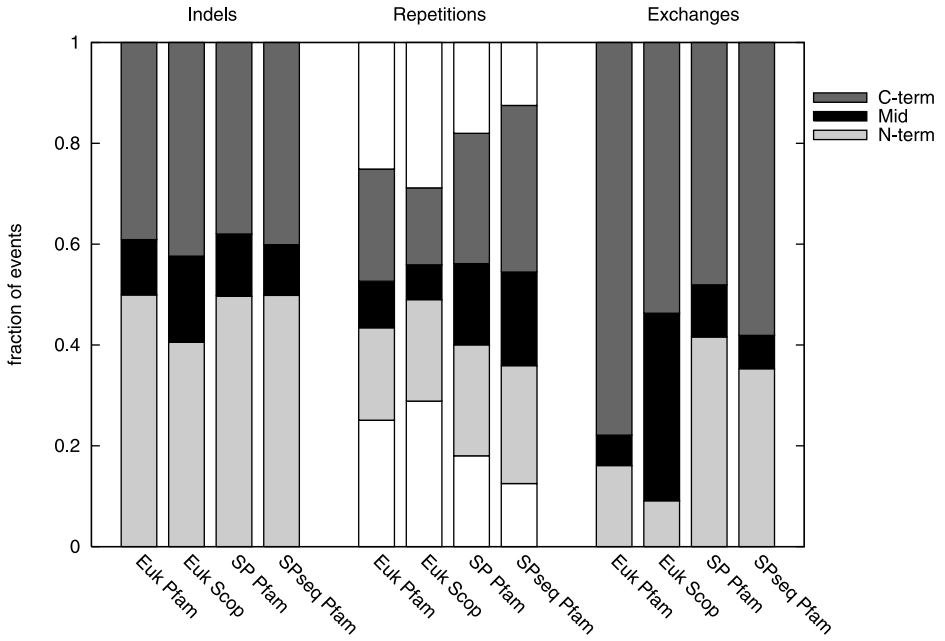

Figure 3. The position of evolutionary events (indels, repetitions and exchanges). Results are given for: eukaryotes using domain distance (Euk) with Pfam and SCOP; SWISS-PROT using domain distance (SP Pfam); and SWISS-PROT using sequence alingments (SPseq Pfam). Only domain architectures with more than two domains were used. For domain repeats where both the query protein and the closest neighbor consist of one domain family, the position of the event cannot be determined. These examples have been divided equally below and above the other bars in the graph (white bars).

dataset and 803 Pfam DAs in the SWISS-PROT set. Indels and repetitions seldom occur in the middle of a protein, i.e. between two domains (see Figure 3). In total, the fraction of events in the middle is 10–13%. A slight preference for indels at the N-terminal can be noted, while the repeats are evenly distributed at both ends. The exchanges, on the other hand, mainly occur at the C terminus. However, as the exchanges are few, the significance of the position of exchanges may be low.

The closest neighbor differs by only one domain for 78–94% of the multi-domain architectures, hence most events can be explained by the addition of a single domain. Insertion/deletion of several domains at a time seems to be rare and most events that involve several domains are found in repetitions (see Table 2). Our results correlate well with previous findings that some frequent domain combinations, or supra-domains, remain intact while new domains are added to the architectures.[6]

To characterize the domain families that have been added in the evolutionary events, the families in the eukaryotic dataset were divided into three groups: families with no events, repeating families and indel families (see Materials and Methods).

Table 2. The number of domains involved in the evolutionary events for: eukaryotes using domain distance (with Pfam and SCOP domains), SWISS-PROT Pfam using domain distance and SWISS-PROT Pfam using sequence alignments (SWISS-PROT seq) both with Pfam domains

Dataset	Event	1D	2D	3D+
Eukaryotes Pfam	Ind	0.91	0.06	0.02
	Rep	0.84	0.09	0.07
	Exc	0.97	0.03	0.00
Eukaryotes SCOP	Ind	0.95	0.04	0.04
	Rep	0.88	0.08	0.04
	Exc	1.00	0.00	0.00
SWISS-PROT Pfam	Ind	0.91	0.08	0.02
	Rep	0.76	0.12	0.12
	Exc	0.99	0.001	0.00
SWISS-PROT seq Pfam	Ind	0.89	0.08	0.03
	Rep	0.72	0.16	0.12
	Exc	0.91	0.09	0.00

The fractions of each event (indels, repetitions and exchanges) involving different number of domains (one domain (1D), two domains (2D) and three or more domains (3D+)).

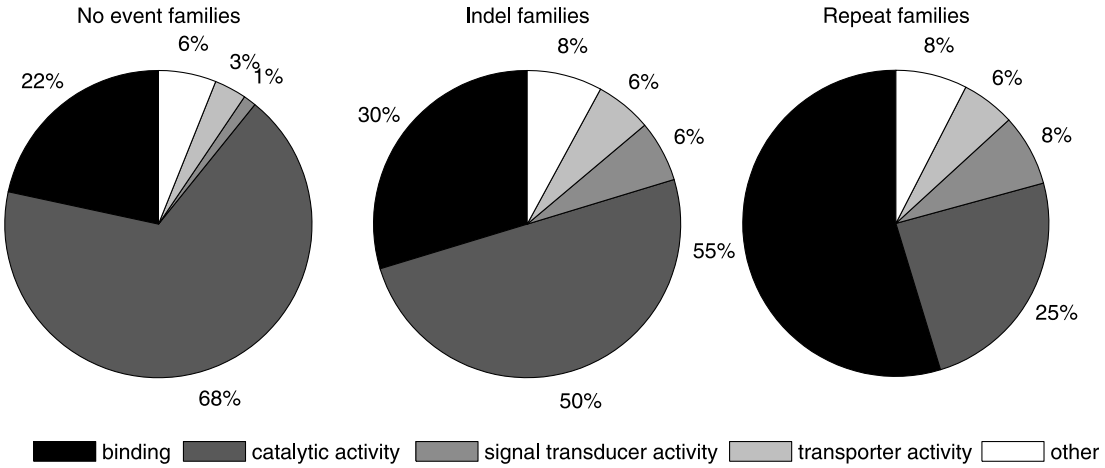

Figure 4. Function of domain families with no events, repeating families and indel families. The fraction of domains with the following GO functions are shown: binding, enzymatic activity, signal transducer activity, transporter activity or other GO function.

It was found that the indel families are, on average, longer than the repeat families (161 ± 7 compared to 103 ± 14 residues). As shown in Figure 4, most of the families with no events are enzymatic, while the added families are often involved in binding, in particular the repeating domain families. This shows that new enzymatic domains seldom are added to an architecture, while new binding domains can alter the enzyme specificity.

As can be expected, the course of evolution is different for different protein classes. We have investigated what events are detected in proteins containing the most common domain families (see Table 3). It is clear that proteins containing repeating domain families, especially C2H2 zinc finger proteins, often evolve by repetition. In protein kinases, on the other hand, mainly indels are found, which is natural, since nearly 70% of them are two or three-domain proteins. Interestingly, the EGF-containing proteins have often evolved by the addition of several domains at a time. The EGF domains are often found in large and complex domain architectures and their insertions are frequently found between domains. Finally, to ensure that no single domain family has biased our data, we excluded DAs containing the ten most common domain families. The results differed less than one percentage point when excluding them separately or all at once.

It is quite possible that the true evolutionary origin was not identified for individual proteins, either due to divergence within the domain family or because it was not present in our dataset. Accordingly, we cannot exclude the possibility that many proteins have evolved by other mechanisms. As mentioned, all calculations are based on the assumption that a domain combination evolved from the domain architecture to which it has the shortest domain distance. While this is a likely scenario, it may not always be the case. When sequence alignments, instead of domain distance, were used to detect neighbors, the frequency of events including multiple domains and exchanges increased, but only marginally. For 80% of the DAs the same closest neighbor is identified with both methods, and most of the cases where the neighbors differ are in proteins with long repeats. In some of the detected exchanges the two domains seem to be alignable and may represent homologous Pfam-A families. Anyhow, the general trends, i.e. that most events contain one single-domain and occur at either the N or C terminus and that exchanges are rare, are also found using sequence similarity to detect the closest neighbors. In addition, the same

Table 3. Summary of domain specific evolutionary events for the five Pfam-A domain families that are found in the largest number of domain architectures (DA) in the eukaryotic dataset

Domain	Name	NoDA	Ind	Rep	1D
PF00096	zf-C2H2	79	0.38	0.62	0.97
PF00069	Pkinase	61	0.82	0.18	0.92
PF00008	EGF	50	0.57	0.41	0.71
PF01391	Collagen	47	0.45	0.47	0.80
PF00560	LRR	43	0.48	0.47	0.85
All		4347	0.59	0.39	0.88

For all DAs that contain the domain family in question, the events were calculated and the number of DAs, the fraction of indels and repetitions and finally the fraction of events including one domain (1D) are shown. The bottom row shows the same statistics with the whole dataset.

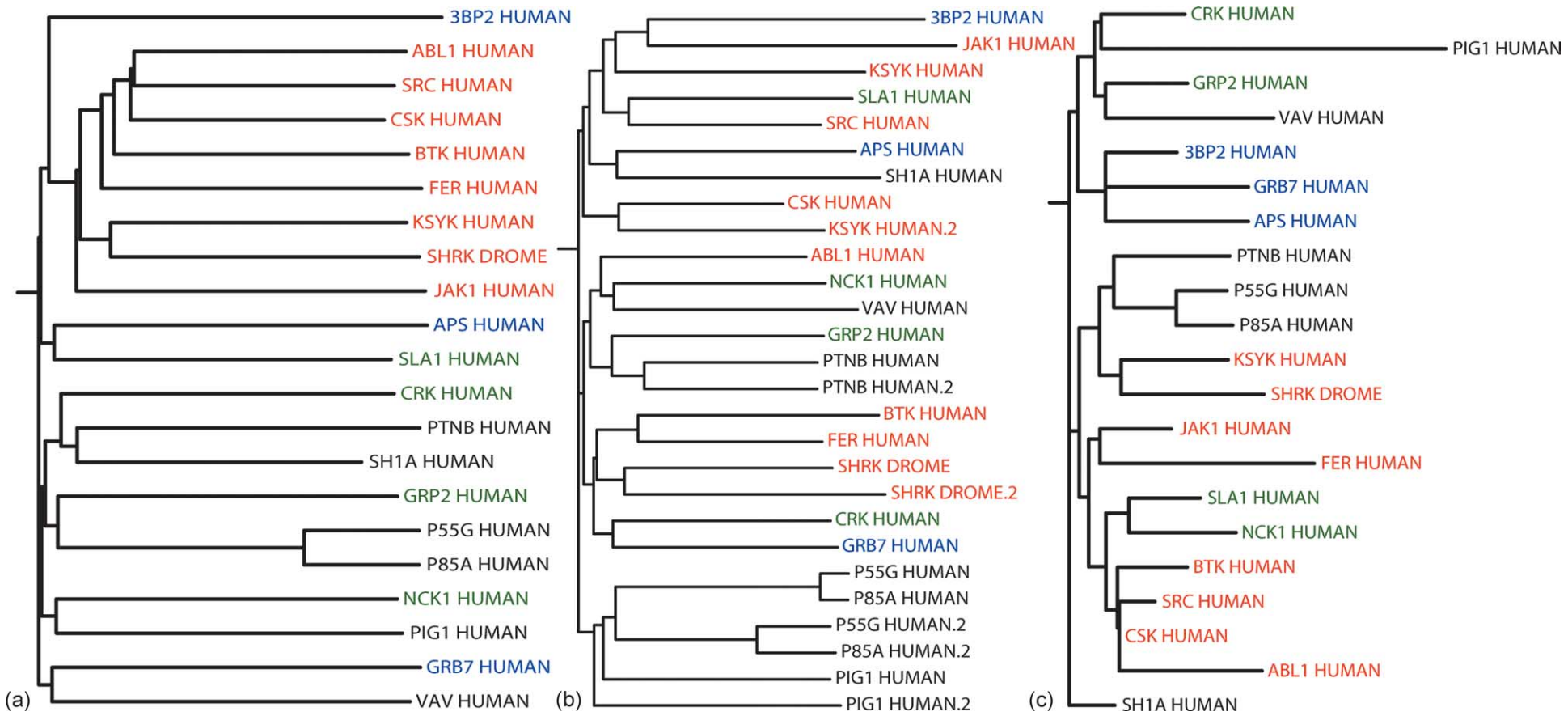

Figure 5. Evolutionary trees for a set of SH2 domain-containing proteins. Trees have been created from sequence alignment using both (a) full-length sequences and (b) with sequences of the SH2 domains. If a protein has two SH2 domains the name of the second domain is indicated with ″2″. (c) A tree has also been created with domain distances. Proteins are colored according to: kinases (red); adaptors with PH-domains (blue); and adaptors with SH3-domains (green). One difference is that the PH adaptors are clustered in the domain distance tree but not in the sequence trees.

trends are seen when SCOP domains instead of Pfam-A domains were used. Therefore, we believe that these trends are not an artifact of the domain distance measure. Accordingly, our observations give an indication of the relative importance of different evolutionary events in the creation of multi-domain proteins. Further, it is likely that different domain recombinations are subjected to different evolutionary constraints. As a consequence, a more detailed study would require a more elaborate definition of domain distances with, for instance, independent weights for different domain families and distinct calculations of domain distance for repeating and non-repeating regions.

Domain Distance Trees in Evolutionary Studies

To obtain an increased understanding of the evolution of multi-domain proteins, the domain distances can be used to build evolutionary trees. Such trees have been created using standard neighbor-joining methods, where each addition/deletion of a domain results in a new branch. Below, we exemplify how such a tree can aid our understanding of the evolutionary events for two large protein families: SH2/PTK (Src homology 2 domain containing protein tyrosine kinases) and the Rho-GEFs (Rho guanine nucleotide exchange factors).

Evolution of SH2-containing non-receptor tyrosine kinases

SH2 is a domain that is important in many signaling pathways for its ability to bind phosphorylated tyrosine residues, and is found in proteins of highly diverse functions, such as tyrosine kinases, phosphatases and adaptor molecules. Although SH2 is not present in prokaryotes (data not shown), a variety of SH2-containing tyrosine kinases have been found in organisms of ancient metazoan phylums, such as sponge, implying that many of the domain rearrangements happened early in metazoan evolution.[20]

For a group of different SH2-containing domain architectures in SWISS-PROT we have created three trees, one based on domain distance, and two based on sequence alignments using ClustalW,[12] with full-length sequences and SH2 domain sequences (see Figure 5). In the protein sequence tree all kinases (Src, Btk, Abl, Fer, Syk and the Jak) are grouped together and a similar organization is seen in the domain distance tree. On the other hand, the pleckstrin homology (PH) containing adaptors are split between different branches in the two sequence trees, whereas they are clustered together in the domain distance tree.

A plausible evolutionary history of the kinases has been reconstructed from the domain distance tree. Their evolution from the smallest common motif (SH2+PTK) most likely started with independent additions of SH3, FCH, SH2 and FERM at the N terminus as described in Figure 6. This example demonstrates how the addition of new domains gives enzymes different specificity and regulation.

Fer, Syk, Shark and Jak all have insertions at the N terminus. Syk has no additional functional domain; however, it has an extra SH2 domain that is used for more specific substrate binding.[21] The Jaks have an insertion of a cytokine receptor-binding FERM domain at the N terminus.[22] In addition, they have a kinase domain inserted at the C-terminal, which is unusual in this family. However, after insertion of the second kinase domain the original function of the first one has been disrupted.[23] While all other insertions seem to have happened early in metazoan evolution, this C-terminal kinase domain is not present in the sponge homologue,[20] and may be the result of a later insertion.

The largest group among these kinases contains an SH3 domain at the N terminus (see lower part of Figure 6). The SH3 domain in the Src-related proteins is, in addition to being important for ligand binding in the active kinase, essential for inactivating the Src kinases.[24,25] The Src-like adaptor protein Slap is closely related to the Src kinases, e.g. Lck, where the SH2 and SH3 domains are involved in negative regulation of T-cell antigen receptor signaling, mediated by Lck.[26,27] The evolutionary relation between the Src kinases and Slap is also demonstrated by the domain sequence similarity (see Figure 5(b)). We propose that Slap evolved from an Src kinase through deletion of the kinase domain, as the SH2 and SH3 domains in the two proteins are closely related.

In summary, the events in the family of non-receptor tyrosine kinases have provided them with regulatory (SH3) or binding function (SH3, PH and FERM) and include mainly N-terminal insertions, except in Abl and Jak, which have C-terminal insertions.

We have studied the evolution of tyrosine kinases using a domain distance tree of SH2 containing proteins. With the tree-creation centered around another domain in the protein family, e.g. PTK or SH3, the evolutionary history of the tyrosine kinases may differ. Individual domains may define the protein family, as SH2 and PTK in the example above. The history of the protein family may then agree with the history of the domains. However, other domains are not determinant for the family and are therefore less suitable for studying its evolution.

Evolution of Rho guanine nucleotide exchange factors

Another group of multi-domain proteins is the family of RhoGEFs. These proteins transduce signals in response to stimulation of membrane-bound receptors, leading to activation of Rho GTPases. All RhoGEFs have a Dbl homology (DH) domain and a pleckstrin homology (PH) domain in

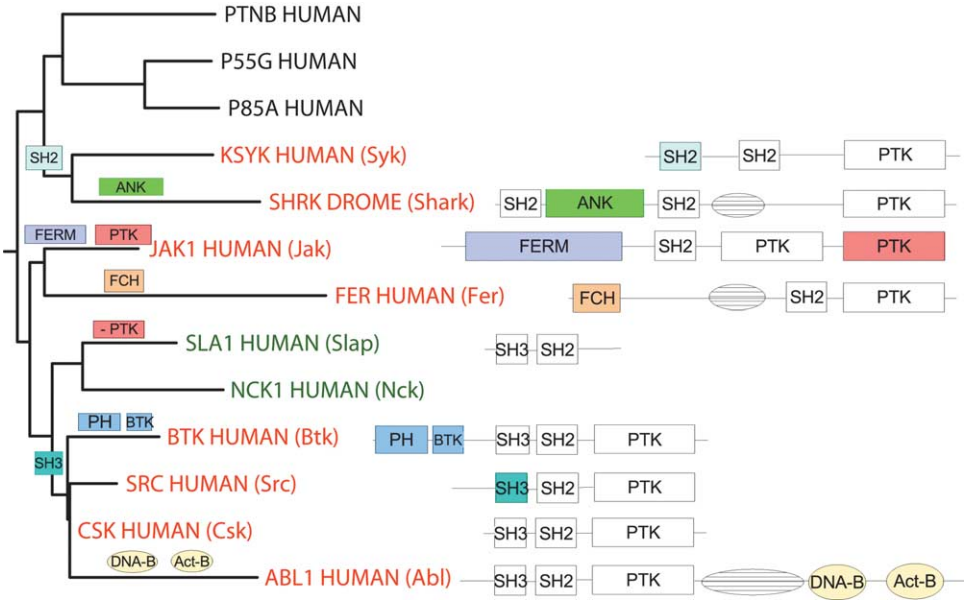

Figure 6. An evolutionary scenario of SH2-containing tyrosine kinases. Pfam-A domain architectures are shown next to the domain distance tree and names are given together with SWISS-PROT identifiers. Ellipses are motifs other than Pfam-A or disordered regions (striped). Starting from a SH2-PTK combination, several different domain architectures have evolved. The additional SH2 domain of Syk and its fly homolog Shark may be either an internal duplication or an insert. Shark also contains an ankyrin repeat between the SH2 domains. *H. vulgaris* has both Syk and Shark homologs, hence the insert of ankyrin must be an ancient event. However, only Syk remains in human while *D. melanogaster* has only Shark and *C. elegans* has neither of them.[40] The Jaks have insertions at both ends: a FERM domain at the N terminus and a second PTK at the C terminus, while Fer evolved from the SH2-PTK ancestor through an insertion of a FCH at the N-terminal, and an additional 200 residues with little structure and no detectable homology to other sequences. Fps, a Fer ortholog in *D. melanogaster*, has an extra insert of 500 partly disordered residues, while the whole N terminus before SH2 is missing in the worm ortholog KI31, possibly indicating a deletion (not shown). An SH3 domain was inserted at the N terminus in Src-related kinases. In Abl a C-terminal tail for DNA and actin binding was then inserted[41] while Btk had a second N terminus insertion of PH and Btk domains. The Btk domain is always found joined with the PH domain, hence it is likely that there has been an insertion with the complete two-domain motif. In Slap the kinase domain has been lost from the Src kinases. PTK, protein tyrosine kinase; FCH, Fes/CIP4 homology domain; PH, pleckstrin homology domain; Ank, ankyrine repeat; DNA-B, DNA-binding motif; Act-B, actin-binding motif.

tandem. The number of RhoGEFs has increased markedly from yeast to human, where 69 distinct members have been found.[28] The evolution of a selection of proteins in this family has been studied here (see Figure 7).

The members of this family contain a large number of different additional domains providing them with a variety of specificities and functionalities. The majority of the events involve insertions of binding domains (e.g. C1, PDZ and SH3), including many plasma membrane targeting domains, but also internal repeats of spectrin and fusions with other multi-domain genes.

The binding domain PDZ of PDZ-RhoGEF, is essential for signaling from Plexin to Rho GTPases by interacting directly with the transmembrane receptor Plexin, leading to regulation of actin rearrangements.[29] Son of sevenless (SOS) and RasGRF1, on the other hand, have fused with RasGEF, resulting in bifunctional proteins with unique functions as coordinators in the pathways of Rho and Ras GTPases, which for SOS has effects on cytoskeletal organization and growth control.[30] SOS also contains an N-terminal insertion of a histone domain, while RasGRF1 has an additional PH domain connected with a calmodulin-binding domain (IQ). Calmodulin seems to be involved in regulation of RasGRF1 activity,[31] whereas the function of the histone domain in SOS is unclear.[32]

Having an SH2 domain connected to a RhoGEF is unique to Vav. Hence, it is not surprising that Vav is the only RhoGEF known to be activated and regulated by tyrosine kinases.[33] Interestingly, Trio is the result of a large number of events starting from Dbl. Through multiple insertions and repetitions it has acquired kinase activity and double GEF domains with different specificities.[34]

In conclusion, the proteins in the RhoGEF family have evolved through many different events and some of the longer domain insertions may be the result of more than one evolutionary event, with the intermediate domain architecture missing from our dataset. There is no bias for either terminus, with six N-terminal and five C-terminal insertions, in contrast with what we saw in the family of SH2/PTK.

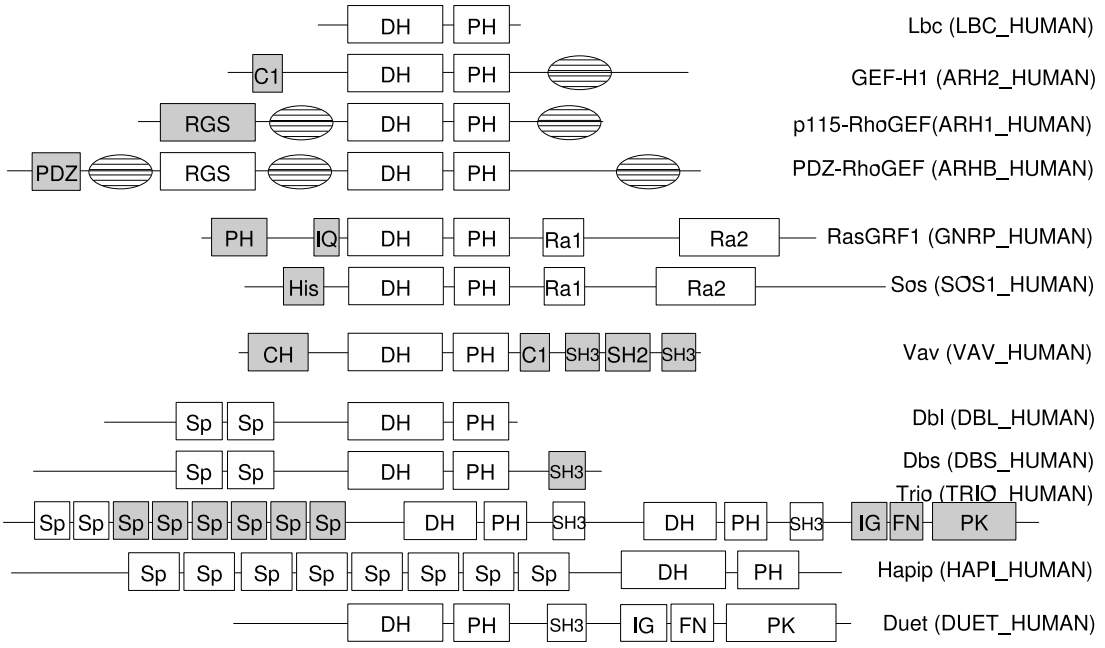

Figure 7. An evolutionary scenario of Rho guanine nucleotide exchange factors (RhoGEF). Pfam-A domain architectures are shown and names are given together with SWISS-PROT identifiers. Shading indicates where domains have been inserted. Striped ellipses are regions of disorder. Starting from the simple structure of Lbc, GEF-H1 has evolved through an insertion of C1 at the N terminus, while RGS was inserted in p115-RhoGEF. An additional insertion of a PDZ domain then followed in PDZ-RhoGEF. Son of sevenless (SOS) and RasGRF1 were both created from a C-terminal fusion with RasGEF. SOS also contains an N-terminal insert of histone, whereas RasGRF1 has an additional PH domain connected with a calmodulin-binding domain (IQ). In Vav, five binding domains have been inserted. How this has happened is unclear, but it is possible that the SH3-SH2-SH3 motif was added simultaneously, perhaps through a fusion with the adaptor molecule Grap. The DH-PH-SH3 motif in Dbs may have been internally duplicated followed by a C-terminal insertion of an IG coupled kinase and tandem duplication of the spectrin repeat resulting in Trio. A split in Trio may then have created Hapip and Duet, two adjacent genes on the chromosome. DH, Dbl-homology; PH, pleckstrin homology; RGS, regulator of G protein signaling; His, histone; Ra1, RasGEF N-terminal; Ra2, RasGEF C-terminal; Sp, spectrin (number of spectrin repeats are in some cases uncertain); IG, immunoglobulin; FN, fibronectin; PK, protein kinase.

Conclusions

We have studied the evolution of multi-domain proteins in terms of domain fusions and repetitions. For each domain architecture, its evolutionary origin was identified based on our novel measure domain distance. Using this measure we have quantified the different evolutionary events leading to complex domain architectures and found that indels are the most common domain events followed by repetitions. The majority of the events can be explained by the addition of single domains. However, in creation of long tandem repeats, several domains are sometimes duplicated simultaneously. Domain families that are not involved in any events mainly perform catalytic functions, while the added families, especially the repeating families, more often are involved in binding.

Protein similarity as measured with domain distance is consistent with both sequence similarity and semantic similarity based on GO annotations, and we believe that domain distances can serve as a complement to traditional methods in evolutionary studies. One such use is creation of evolutionary trees based on domain distances. Studies of SH2-containing kinases and RhoGEFs provide two examples of the domain rearrangements within protein families, where many single domains have been inserted at the N or C terminus, resulting in enzymes with new specificities and regulatory functions.

Materials and Methods

Protein set

Two datasets were used for calculation of evolutionary events. The first dataset was SWISS-PROT release 44 (5 July 2004)[35] with 153,871 proteins. The Pfam-A[36] and Pfam-B domain assignments were found in SwissPfam†.

The other dataset consisted of proteins from seven eukaryotic proteomes (*Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*). In the case of multiple splice forms, only the longest transcript was kept. Both Pfam-A and SCOP[37] domains were assigned using HMMER, for details see our

† ftp://ftp.sanger.ac.uk/pub/databases/Pfam/.

Table 4. Distribution of domain architectures (DAs) in each dataset

Dataset	Proteins	noDA	1D	2D	3D	4D	5-9D	10D+
Eukaryotes Pfam	39,861	4347	0.571	0.208	0.077	0.040	0.067	0.036
Eukaryotes SCOP	35,726	2596	0.285	0.723	0.158	0.059	0.061	0.014
SWISS-PROT Pfam	80,729	5747	0.707	0.153	0.061	0.024	0.041	0.014

The total number of DAs (noDA) followed by the fraction of single-domain architectures (1D), two-domain architectures (2D), etc. The three datasets are eukaryotes assigned with Pfam-A or SCOP and SWISS-PROT assigned with Pfam-A.

previous study.[2] Only completely assigned proteins, i.e. with no more than 100 consecutive unassigned residues, were used. For more details about the number of domain architectures in each dataset, see Table 4.

Domain distances

The domain distance between two proteins is calculated from the alignment of the "domain sequences", i.e. each domain family in the DA can be seen as one symbol in the sequence. A simple local dynamic programming was applied to create the alignment using an identity matrix, i.e. a match is given score 1, gap openings −0.01 and gap extensions −0.001, while mismatches are not punished. Other alignment parameters were also tried on a limited scale and it was found that the inclusion of small gap-penalties was necessary to avoid sporadic gaps in long repeating regions. If a mismatch penalty was used, the number of "exchange" events decreased, as it then is more favorable to create two gaps instead of aligning the domains. The domain distance is calculated as the number of unmatched domains in the alignment (see Figure 1). Two proteins with equal domain architectures will have a domain distance zero and, on the opposite, proteins with no domains in common are considered unrelated, and the domain distance is infinite.

Functional comparisons

Functional comparisons were performed with GOGraph,[17] which calculates semantic similarity between two proteins based on the combination of all their gene ontology (GO) annotations with traceable author statement (TAS). GOgraph handles relations between terms, differences in specificity level and the information content of the terms. One problem with the GO terms, however, is that many proteins with similar functions and equal domain-architectures do not have the same GO annotations, since they may have been annotated at different specificity levels or the assignments may be partial.

In the dataset used for calculation of functional similarity we used only proteins with GO annotations and complete domain assignments. Only one representative of proteins with identical domain architectures and GO annotations was included. This resulted in a dataset with 11,708 proteins. TAS GO annotations were found for 5214 of these proteins in our non-redundant SWISS-PROT dataset, of which 3874 had annotated molecular function, 4099 biological process and 2779 cellular compartment.

For each of the three classes of GO annotations, the semantic similarity between all pairs of proteins with at least one domain in common was calculated using GOGraph. In total, semantic similarity was calculated for 66,262 protein pairs (45,322 with molecular function, 53,101 with biological process and 26,022 with cellular component). Semantic similarity was also calculated for a subset of 2000 randomly selected protein pairs with no domains in common (infinite domain distance) for comparison.

Sequence similarity

Sequence similarity was measured with Blast[38] bit scores, either with the whole protein sequences or between the sequences of domains from the same family. Sequence similarity was compared to domain distances for the same data set that was used in comparisons with semantic similarity.

Extracting statistics on evolutionary events

The frequencies of different evolutionary events were calculated on two datasets (see above). First, domain distances were calculated for a query DA to all other DAs of equal or shorter length, as we assume that evolution progresses towards more complex DAs (see Figure 1(b)). The query DA was assumed to have originated from the DA to which it had the shortest domain distance. If several DAs had the same shortest distance, statistics for all of them was extracted and the average was calculated. For DAs that do not share domains with any other DA no events were counted. The query and the neighbor DAs were aligned and the number of evolutionary events, i.e. indels, repeats and exchanges were counted, as well as the number of involved domains and the position of the event, i.e. N-terminal (before first domain), middle (between domains) or C-terminal (after last domain).

For comparison, closest neighbors were also identified using Blast[38] bit scores for the SWISS-PROT data set. For some of the proteins, no Blast match could be found. Many DAs are over-represented in SWISS-PROT and all identical DAs, with the same closest neighbor, were pooled together and counted once.

GO function of domain families

For all proteins in the SWISS-PROT data set with GO annotations and Pfam-A domains, the GO annotation at the highest level of molecular function was extracted (antioxidant activity, binding, catalytic activity, chaperone activity, chaperone regulator activity, enzyme regulator activity, molecular function unknown, motor activity, nutrient reservoir activity, obsolete molecular function, signal transducer activity, structural molecule activity, transcription regulator activity, translation regulator activity, transporter activity and triplet codon-amino acid adaptor activity). Subsequently, the GO function most commonly associated with a domain family was identified. However, the GO function of domain families is based on the annotations of the proteins they are found in, e.g. a non-enzymatic domain mainly found in enzymes is wrongly annotated as enzymatic.

The families in the eukaryotic dataset were divided into

three groups: families with over 30% of the occurrences involved in an indel, families with over 30% of the occurrences involved in a repeat event, and families that never are involved in any events. Only events in domain combinations with three or more domains were considered, as the inserted domain cannot be determined for two-domain architectures. For each of these groups, the frequencies of the different GO functions were calculated.

Tree creation

Trees were created for a subset of SH2-containing proteins from SWISS-PROT. Proteins with only the SH2 domain in common with the other proteins in the set were excluded and for each DA we selected one sequence from the human genome. Only Shark (SHRK_DROME) was chosen from another organism (*D. melanogaster*) because it has a distinct DA, which has been lost from the human genome. Domain distances were calculated for Pfam-A/Pfam-B DAs and trees were created with Phylip 3.6† using neighbor joining. Sequence alignment trees were created with ClustalW,[12] both for full-length sequences and for all SH2 domain sequences in these proteins.

Acknowledgements

This work was supported by grants from the Swedish Natural Sciences Research Council, and a STREP grant from European Union FP6 program *via* the GeneFun project, project number 503567.

References

1. Apic, G., Gough, J. & Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325.
2. Ekman, D., Björklund, Å. K., Frey-Skött, J. & Elofsson, A. (2005). Multi-domain proteins in the three kingdoms of life-orphan domains and other unassigned regions. *J. Mol. Biol.* **348**, 231–243.
3. Gerstein, M. & Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci.* **7**, 445–456.
4. Liu, J. & Rost, B. (2004). Chop proteins into structural domain-like fragments. *Proteins: Struct. Funct. Bioinf.* **55**, 678–688.
5. Qian, J., Luscombe, N. M. & Gerstein, M. (2001). Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* **313**, 673–681.
6. Vogel, C., Teichmann, S. & Pereira-Leal, J. (2005). The relationship between domain duplication and recombination. *J. Mol. Biol.* **346**, 355–365.
7. Vogel, C., Berzuini, C., Bashton, M., Gough, J. & Teichmann, S. A. (2004). Supra-domains: evolutionary units larger than single protein domains. *J. Mol. Biol.* **336**, 809–823.
8. Liu, M., Walch, H., Wu, S. & Grigoriev, A. (2005). Significant expansion of exon-bordering protein domains during animal proteome evolution. *Nucl. Acids Res.* **33**, 95–105.
9. Hegyi, H. & Gerstein, M. (2001). Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* **11**, 1632–1640.
10. Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425.
11. Todd, A., Orengo, C. & Thornton, J. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
12. Thompson, J. D., Higgins, D. & Gibson, T. (1994). ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
13. Geer, L., Domrachev, M., Lipman, D. & Bryant, S. (2002). Cdart: protein homology by domain architecture. *Genome Res.* **12**, 1619–1623.
14. Storm, C. E. & Sonnhammer, E. (2001). Nifas: visual analysis of domain evolution in proteins. *Bioinformatics*, **17**, 343–348.
15. Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C. & Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14**, 208–216.
16. Marcotte, E., Pellegrini, M., Yeates, T. O. & Eisenberg, D. (1999). A census of protein repeats. *J. Mol. Biol.* **293**, 151–160.
17. Lord, P. W., Stevens, R. D., Brass, A. & Goble, C. A. (2003). Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.
18. Snel, B., Bork, P. & Huynen, M. (2000). Genome evolution. Gene fusion *versus* gene fission. *Trends Genet.* **16**, 9–11.
19. Kummerfeld, S. & Teichmann, S. (2005). Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet.* **21**, 25–30.
20. Suga, H., Katoh, K. & Miyata, T. (2001). Sponge homologs of vertebrate protein tyrosine kinases and frequent domain shufflings in the early evolution of animals before the parazoan-eumetazoan split. *Gene*, **280**, 195–201.
21. Ottinger, E., Botfield, M. & Shoelson, S. (1998). Tandem sh2 domains confer high specificity in tyrosine kinase signaling. *J. Biol. Chem.* **273**, 729–735.
22. Hilkens, W., Is'harc, H., Lillemeier, B., Strobl, B., Bates, P., Behrmann, I. & Kerr, I. (2001). A region encompassing the ferm domain of jak1 is necessary for binding to the cytokine receptor gp130. *FEBS Letters*, **505**, 87–91.
23. Wilks, A., Harpur, A., Kurban, R., Ralph, S., Zurcher, G. & Ziemiecki, A. (1991). Two novel protein-tyrosine kinases, each with a second phosphotransferase-related catalytic domain, define a new class of protein kinase. *Mol. Cell. Biol.* **11**, 2057–2065.
24. Xu, W., Harrison, S. & Eck, M. (1997). Three-dimensional structure of the tyrosine kinase c-src. *Nature*, **385**, 595–602.
25. Sicheri, F., Moarefi, I. & Kuriyan, J. (1997). Crystal structure of the src family tyrosine kinase hck. *Nature*, **385**, 602–609.
26. Sosinowski, T., Pandey, A., Dixit, V. & Weiss, A. (2000). Src-like adaptor protein (slap) is a negative regulator of t cell receptor signaling. *J. Expt. Med.* **191**, 463–474.
27. Tang, J., Sawasdikosol, S., Chang, J. & Burakoff, S.

† http://evolution.genetics.washington.edu/phylip.html.

(1999). Slap, a dimeric adapter protein, plays a functional role in T cell receptor signaling. *Immunology*, **96**, 9775–9780.
28. Rossman, K., Der, C. & Sondek, J. (2005). Gef means go: turning on rho gtpases with guanine nucleotide-exchange factors. *Nature Rev. Mol. Cell Biol.* **6**, 167–180.
29. Driessens, M., Olivo, C., Nagata, K., Inagaki, M. & Collard, J. (2002). B plexins activate rho through pdz-rhogef. *FEBS Letters*, **529**, 168–172.
30. Nimnual, A. S., Yatsula, B. A. & Bar-Sagi, D. (1998). Coupling of Ras and Rac guanosine triphosphatases through the Ras exchanger Sos. *Science*, **279**, 560–563.
31. Baouz, S., Jacquet, E., Bernardi, A. & Parmeggiani, A. (1997). The N-terminal moiety of cdc25(mm), a gdp/gtp exchange factor of ras proteins, controls the activity of the catalytic domain. Modulation by calmodulin and calpain. *J. Biol. Chem.* **272**, 6671–6676.
32. Sondermann, H., Soisson, S., Bar-Sagi, D. & Kuriyan, J. (2003). Tandem histone folds in the structure of the N-terminal segment of the ras activator son of sevenless. *Structure*, **11**, 1583–1593.
33. Bustelo, X. (2000). Regulatory and signaling properties of the vav family. *Mol. Cell. Biol.* **20**, 1461–1477.
34. Debant, A., Serra-Pages, C., Seipel, K., O'Brien, S., Tang, M., Park, S.-H. & Streuli, M. (1996). The multidomain protein Trio binds the LAR transmembrane tyrosine phosphatase, contains a protein kinase domain, and has separate rac-specific and rho-specific guanine nucleotide exchange factor domains. *Proc. Natl Acad. Sci. USA*, **93**, 5466–5471.
35. Bairoch, A. & Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement trembl. *Nucl. Acids Res.* **24**, 17–21.
36. Sonnhammer, E., Eddy, S. & Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Struct. Funct. Genet.* **28**, 405–420.
37. Murzin, A., Brenner, S., Hubbard, T. & Chothia, C. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
38. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
39. Aroul-Selvam, R., Hubbard, T. & Sasidharan, R. (2004). Domain insertions in protein structures. *J. Mol. Biol.* **338**, 633–641.
40. Steele, R., Stover, N. & Sakaguchi, M. (1999). Appearance and disappearance of syk family protein-tyrosine kinase genes during metazoan evolution. *Gene*, **239**, 91–97.
41. McWhirter, J. & Wang, J. (1993). An actin-binding function contributes to transformation by the bcr-abl oncoprotein of philadelphia chromosome-positive human leukemias. *EMBO J.* **12**, 1533–1546.

Edited by M. Levitt

(Received 19 May 2005; received in revised form 19 August 2005; accepted 26 August 2005)
Available online 21 September 2005