

Evolution of Multidomain Proteins

Relevant Literature

- *C. Chothia, J. Gough, C. Vogel, S. A. Teichmann, "Evolution of the Protein Repertoire", www.sciencemag.com, Science VOL 300, 13 June 2003*
- *T. Przytycka, G. Davis, N. Song, D. Durand, "Graph Theoretical Insights into Evolution of Multidomain Proteins", RECOMB 2005, LNBI 3500, pp. 311-325, 2005*

Table of Contents

1. Proteins and their functions	1
1.1 Protein Domains	2
1.2 Domain Family	2
2. Increase in the protein repertoire	3
3. Analysis of Evolution	4
3.1. Protein Repertoire	4
3.2. Domain Combinations	5
3.3. Supra Domains	6
4. Metabolic Pathway Formation	7
5. Multidomain Protein Mystery	7
6. Protein Family Analysis	8
7. Parsimony Model	8
7.1. Dollo Parsimony	9
7.2. Digression: maximum parsimony example	9
8. Evolution of Multidomain Proteins	11
8.1. Protein Overlap Graph	12
8.2. Domain Overlap Graph	12
8.3. Static Dollo Parsimony	12
8.4. Conservative Dollo Parsimony	13
8.5. Motivation of Using Parsimony	13
9. Analyzing the Graph	13
10. Experimental Results	15

1. Proteins and their functions

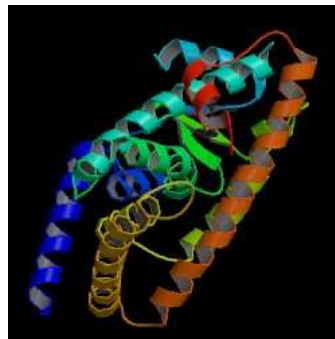
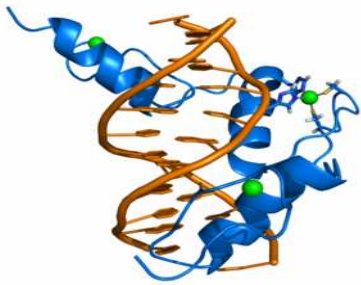
These are large Organic Compounds made of amino acids which fold into specific Structures, unique to each protein. These proteins are functionally extremely important. Proteins are chief actors in the cell. They bind to other molecules specifically and tightly

at the binding site. Besides, they act as enzymes to catalyze chemical reactions. Antibodies are proteins that bind to antigen and target them for destruction

1.1 Protein Domains

They are primary constituent of proteins. It is a conserved evolutionary structural unit:

1. Assumed to fold independently
 2. Observed in different proteins in the context of different neighboring domains
 3. Whose coding sequence can be duplicated and/or undergo recombination
- Small proteins contain just one domain while larger proteins are formed by combination of domains. These domains are generally endowed with a specific functionality. Hence they play a very important role functionally. Example of such a domain can be the binding domain. The length of the domains can range arbitrarily but in general fall between 100 to 250 nucleotides long.



- a. protein Zif268 (blue) containing three **zinc fingers** domains in complex with DNA (orange). The coordinating amino acid residues of the middle zinc ion (green) are highlighted.
- b. Binding Domain

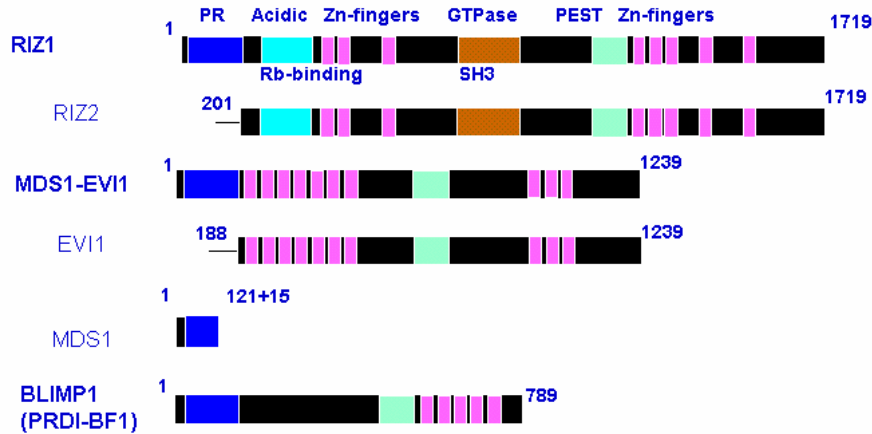
1.2 Domain Family

A domain family is a collection of small proteins and/or parts of larger ones that descend from a common ancestor.

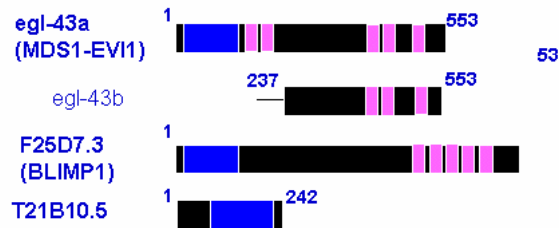
The following diagram depicts the PR domain family members:

PR (*PRDI-BF1* and *RIZ* homology) domain family

HUMAN



C. ELEGANS



It is very important for us to know the domain family relationships. However there are several hindrances towards this end. Namely, it is kind of difficult to discover distant relationships between protein families. A prior knowledge of the protein 3-D structure does help. We only know family relationships and domain structures of proteins of known structures or proteins homologous to proteins of known structures.

This leads to the important fact that the numbers of members in a protein family follow a Pareto distribution. This idea is sometimes expressed more simply as the Pareto principle or the "80-20 rule" which says that 20% of the population owns 80% of the wealth. Thus in this case it simply states that few families have many members whereas many families have few members. The basic intuition being the functional aspects of few families. Since their properties molecular functions, they tend to have more members.

2. Increase in the protein repertoire

There are several actions going on which results in the increase in the protein repertoire. This includes the duplication of coding sequences for one or more domain, divergence of

duplicated sequences by mutations, deletions and insertions producing modified structures that may have useful new properties and recombination of genes that results in new arrangements of domains. These mechanisms have long been believed to be the source of new proteins, and rates at which they occur have been calculated recently. The new findings discussed here come from the use of structural information to analyze genome sequences. This provides for the first time a quantitative view of the nature and extent of these processes. The authors here investigate into the issue of progressive increment in protein repertoire. They try to analyze and explain this phenomenon by looking at the protein structure of known homologous proteins.

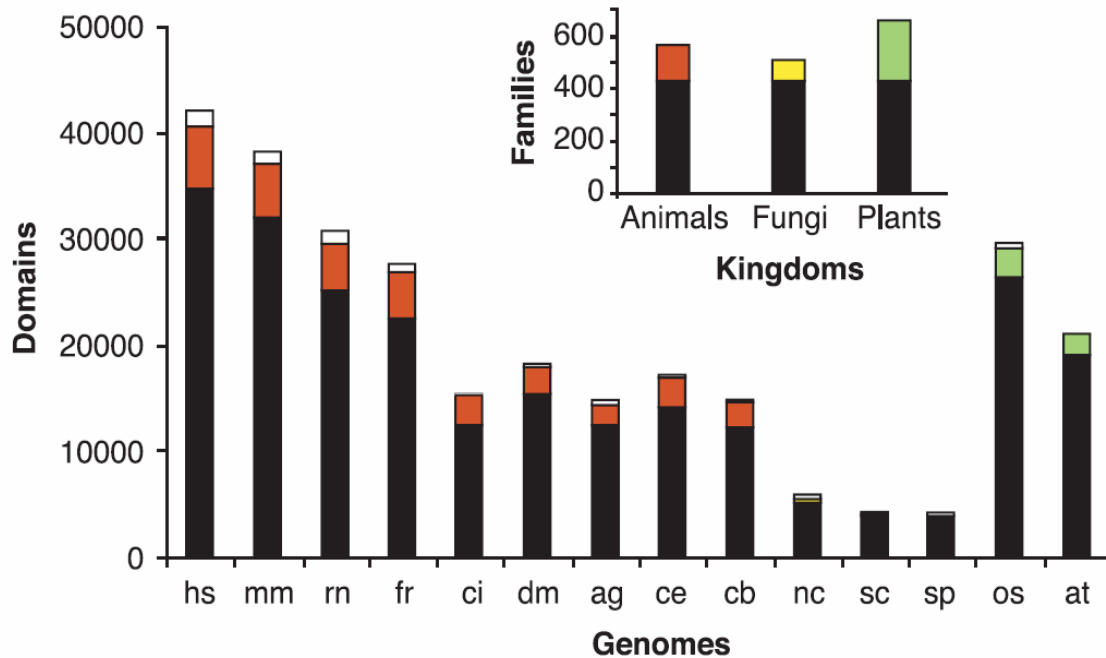
3. Analysis of Evolution

Most proteins have been formed by gene duplication, recombination, and divergence. 50% of sequences in the currently known genomes homologous to proteins of known structure, and these data provide a quantitative description and can suggest hypotheses about the origins of these processes. The evolutionary relationships of domains in proteins of known structure are described in the Structural Classification of Proteins (SCOP) database. This information can be used to infer the family relations of the domains in the genome sequences that are homologous to proteins of known structure. In other words, nearly all proteins have structural similarities with other proteins and, in some of these cases, share a common evolutionary origin. The SCOP database, created by manual inspection and abetted by a battery of automated methods, aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification.

In vertebrates, these domains belong to one of 750 different families in each genome, and the average number of domains in a family is close to 50; those in invertebrates come from one of 670 families in each genome, and the average family size is close to 20. Plants have a similar range of values. Domains in yeast and bacteria with large genomes come from 550 families, and those in small parasitic bacteria come from 220 families. The average size of the known protein families in these two groups is about eight and two, respectively.

3.1. Protein Repertoire

Generally the common families form the bulk of the protein repertoire. They also have greater contributions. The diagram below gives the same idea.



As expressed before the diagram shows that 50% of eukaryote sequences are homologous to domains in proteins of known structure. The numbers of domains that belong to the 429 families common to all 14 eukaryotes studied are shown in black. Additional contributions of families common to the genomes in only one kingdom are shown in red for animals, in yellow for fungi, and in green for plants. For the animal genomes—human (hs), mouse (mm), rat (rn), puffer fish (fr), sea squirt (ci), fruit fly (dm), mosquito (ag), and nematodes (ce and cb)—there are 136 additional common families. For the three fungi—bread mold (nc), budding yeast (sc), and fission yeast (sp)—there are 75 additional common families. For the two plants—rice (os) and cress (at)—there are 229 additional common families.

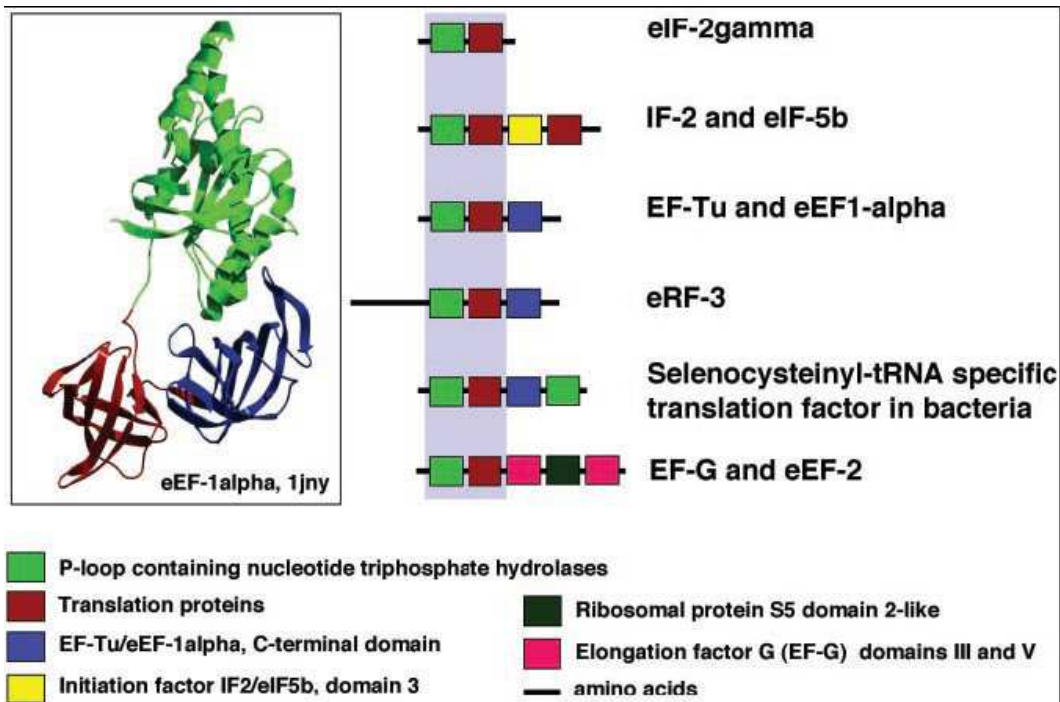
3.2 Domain Combinations

There has been evidence that often 2 or more domains combine to form newer proteins. These also reveal that domains of some families can combine with domains of several other families. The advent of complete genome sequences made it possible to study the extent to which this occurs. Rough estimates indicate that two-thirds of prokaryote proteins have two or more domains. In eukaryotes, in which recombination is even more common, about four-fifths of proteins are multidomain. The tendency of eukaryote

proteins to have more domains than their prokaryote homologs (that is, more complex architectures and properties) has been termed “domain accretion”. There are some known combinations. We know around 1100 protein families. Considering all possible combinations, we find that the maximum such combinations is to the order of $1100^2 = 1210000$ different pair wise combinations. However studies showed that only 2500 pairwise combinations were found in 85 different genomes. The reasoning behind is not all such pairwise combinations are useful. Hence only the important ones exist in nature. Thus only a small subset is stable.

3.3 Supradomains

Domains can be shuffled by recombination to create proteins with new arrangements of domains. We found two-domain and three-domain combinations that recur in different protein contexts with different partner domains. The domains within these combinations have a particular functional and spatial relationship. These units are larger than individual domains and we term them "supra-domains". These are in general larger than the individual domains. Over one-third of all structurally assigned multi-domain proteins contain these over-represented supra-domains. This means that investigation of the structural and functional relationships of the domains forming these popular combinations would be particularly useful for an understanding of multi-domain protein function and evolution as well as for genome annotation. These and other supra-domains were analysed for their versatility, duplication, their distribution across the three kingdoms of life and their functional classes.



The diagram above depicts an example of a supradomain. The P-loop-containing NTP hydrolase domain and the Translation Proteins domain occur in prokaryotic and eukaryotic translation factors that hydrolyze guanosine triphosphate (GTP). GTP hydrolysis in the P-loop domain drives the conformational change in the Translation Proteins domain, which is then transmitted onto the ribosome. The supradomain occurs in 35 different domain architectures, and 6 of these are given here. The inset at left shows a protein of known structure, which contains the supradomain. IF, initiation factor; EF, elongation factor; RF, release factor; tRNA, transfer RNA.

4. Metabolic Pathway Formation

A **metabolic pathway** is a series of chemical reactions occurring within a cell, catalyzed by enzymes, resulting in either the formation of a **metabolic product** to be used or stored by the cell, or the initiation of another metabolic pathway (then called a *flux generating step*). Many pathways are elaborate, and involve a step by step modification of the initial substance to shape it into the product with the exact chemical structure desired.

Various metabolic pathways within a cell form the cell's metabolic network. The metabolic pathway a substrate enters depends on the needs of the cell, i.e. the specific combination of concentrations of the anabolical and catabolical end products (the energetics of the flux-generating step). Metabolic pathways include the principal chemical, mostly enzyme-dependent, reactions that an organism needs to keep its homeostasis.

This brings us to the main problem statement:

How does the duplication, divergence and recombination process of the proteins fit into the formation or extension of pathways?

There are 2 possible solutions:

1. The first proposes that, because substrates in a pathway retain some similarities in structure, the enzymes within a pathway could evolve by gene duplications and a divergence in which their catalytic mechanisms were changed and some aspects of their recognition properties retained.
2. The second proposed that enzymes are recruited across pathways, with the duplicated enzymes conserving their catalytic function but evolving different substrate specificities

5. MultiDomain Protein Mystery

There are 3 specific questions:

1. Are new domains acquired infrequently or often enough that the same combinations of domains will be repeated through independent events? Thus it investigates whether there are known patterns in the combination of domains.



The diagrams above depict 3 sets of domains formed via different combination of domains up the tree. We see that in 2 sets of domains are common in the above 2 trees.

2. Once domain architectures are created, do they persist?
3. If the domain is present in ancestral proteins, is it likely to observe them in current proteins?

6. Protein Family Analysis

Traditionally, protein family was modeled via trees using sequence alignment information. However this method did not scale well as there are situations where there are heterogeneous domains involved. In such cases the modeling gets difficult and often infeasible to achieve. So the proposed solution exploits graph structures to study multi domain protein evolution.

7. Parsimony model

A **phylogenetic tree**, also called an **evolutionary tree** or a **tree of life**, is a [tree](#) showing the [evolutionary](#) interrelationships among various [species](#) or other entities that are believed to have a [common ancestor](#). A *phylogenetic tree* is a form of a [cladogram](#). In a phylogenetic tree, each node with descendants represents the most recent common ancestor of the descendants, and edge lengths correspond to [time](#) estimates. In the parsimony model we assume the presence of a phylogenetic tree with each node

described by a set of characters, one for each domain. If a particular domain is present in the node we represent it as 1 else as 0.

1 → domain is present in node

0 → domain is absent in node

We will also define the concept of state change. If the state of a node changes from 0 → 1 i.e. the domain is present in the node from being absent, we say it had a *gain*. Conversely, a state change of 1 → 0 is termed as a *loss*.

- State Change:
 - **0 → 1**: Gain
 - **1 → 0**: Loss
- Perfect phylogeny: This is defined as the phenomenon that each character state change occurs at most 1 time.

7.1 Dollo Parsimony

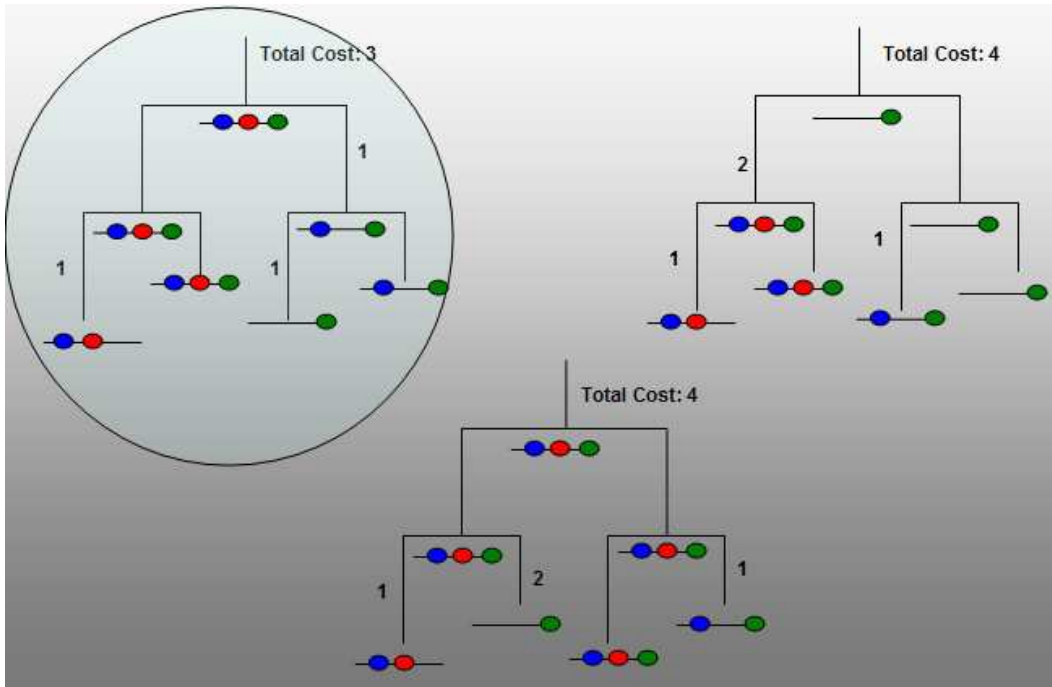
- A character may change state from zero to one *only* once, but from one to zero *multiple* times. In other words gaining is allowed just once while losing is not bounded.
- Appropriate for complex characters that are **hard to gain** but relatively **easy to lose**

7.2 Digression: maximum parsimony example

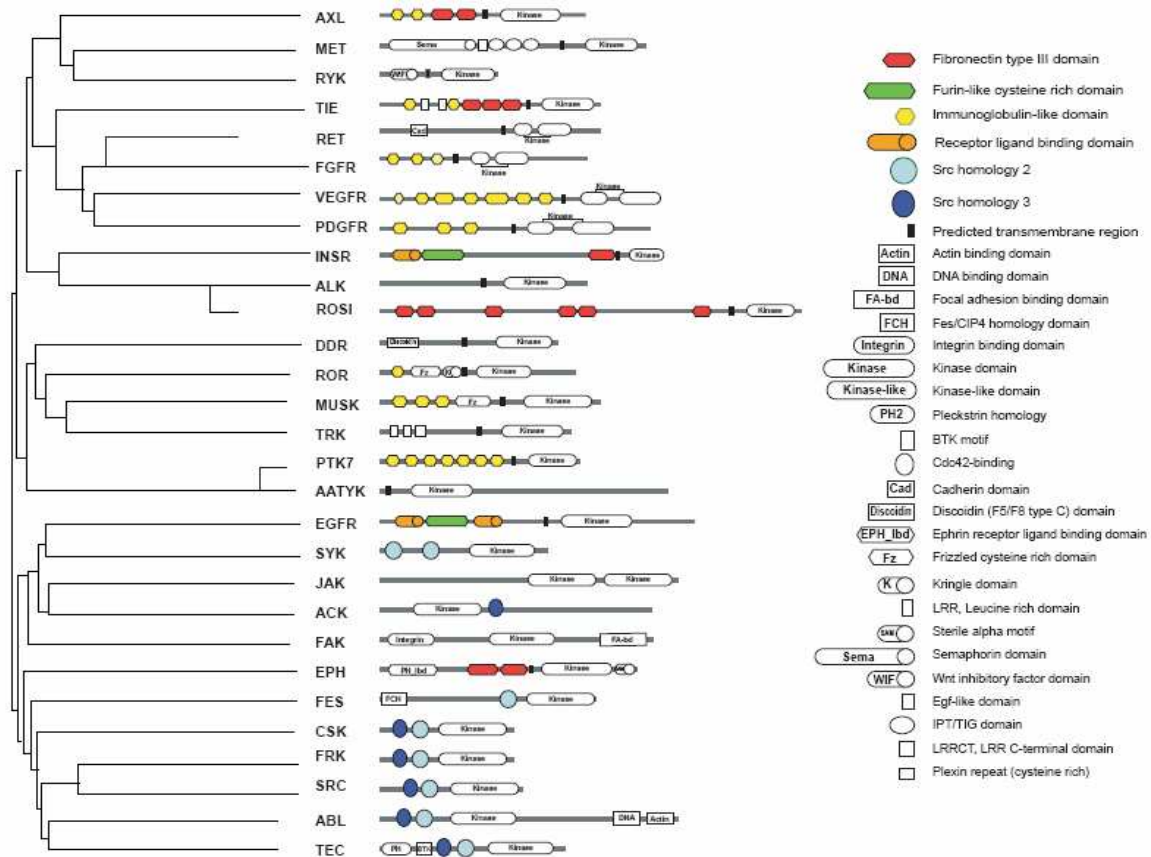
Starting from 3 domains namely D1,D2,D3 we have to find the tree that attains 4 sequences with minimum number of evolutionary changes. Stating it more formally, we want to find a model such that we minimize the total number of insertions and deletions

D1	D2	D3	
●	●	●	
1	1	0	
1	1	1	
0	0	1	
1	0	1	

Here the 4 sequences are shown in the right most column. What the algorithm basically does is try to generate all possible trees that achieve the causes the domains to evolve into the 4 sequences and then select the tree with the minimum cost. Cost here is measured by the number of insertions or deletions. Some graphs typically generated by such an algorithm are as follows:



Clearly the algorithm will choose the first tree as it has the minimum cost of 3.



This figure is of a phylogenetic tree of family protein tyrosine kinase family, constructed from an Multiple Sequence Alignment (MSA) of the kinase domain. It is interesting to notice that the tree is not optimal with respect to a parsimony criterion minimizing the total number of insertions and deletions. For example, if architectures INSR and EGFR were siblings (the only two architectures containing the Furin-like cysteine rich and Receptor ligand binding domains) the number of insertions and deletions would be smaller.

8. Evolution of multi domain proteins

There are several ways of formation of multi domain proteins:

1. Gene Fusion: The accidental joining of DNA of two genes, such as can occur in a translocation or inversion. Gene fusions can give rise to hybrid proteins or to the misregulation of the transcription of one gene by the cis regulatory elements (enhancers) of another.
2. Domain Shuffling: Rearrangement of segments of one or more genes, each segment coding for a structural domain in the gene product, to create a new gene.
3. Retrotransposition of Exons

In the current representation we try to realize each of them via 2 specific methods:

1. Domain Merge: Any process that unites two or more previously separate domains in a single protein



2. Domain Deletion: Any process in which a protein loses one or more domains



8.1. Protein Overlap Graph

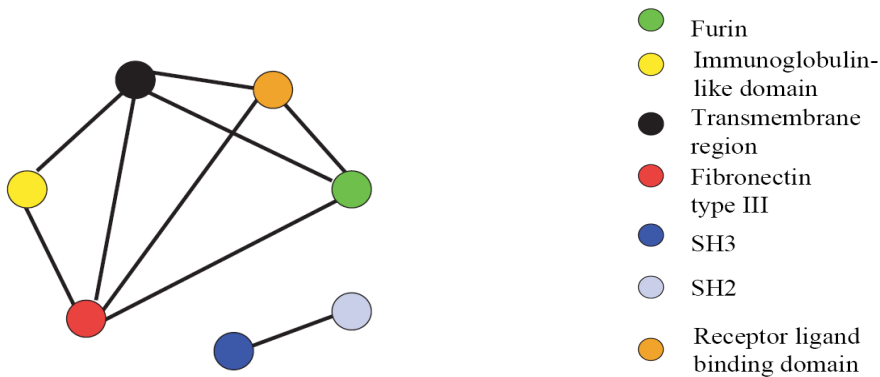
Vertices : proteins

Edges: present if 2 proteins (nodes/vertices) share a domain

8.2 Domain Overlap Graph

Vertices: protein domains

Edges: present between 2 nodes if there exists a protein connecting the 2 domains represented by the 2 nodes respectively.



Thus here for example, Furin and Fibronectin type III share a protein.

8.3 Static Dollo Parsimony

- For any ancestral node, the set of characters in state one in this node is a subset of the set of character in state one in some leaf node. We assume here that more than one character can change in one step.
- This is consistent with a history in which no ancestor contains a domain not seen in a leaf node. Thus it is a restricted form of dollo parsimony.

8.4 Conservative Dollo Parsimony

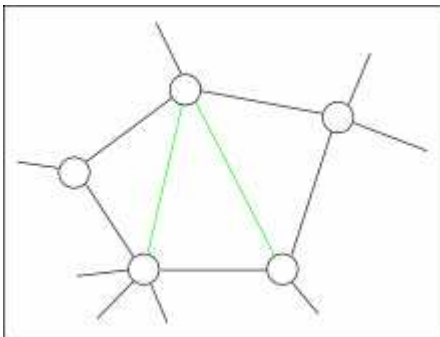
- For any ancestral node and any pair of characters that appear in state one in this node, there exists a leaf node where these two characters are also in state one
- Consistent with a history in which every instance of a domain pair came from a single merge event
- If domains acting in concert offer a selective advantage, it is unlikely that the pair once formed would later separate

8.5 Motivation for using parsimony

For a particular protein family we try to find the existence of conservative dollo parsimony. However if such existence is not found, we can infer that either Single Insertion Assumption is false or Conservative Assumption is too strong. Thus non-existence of conservative Dollo parsimony provides a proof that at least one of these two assumptions is incorrect. On the other hand, existence of conservative Dollo tree does not provide a proof of correctness of the model but only evidence that the assumptions are consistent with the data.

9. Analyzing the Graph

First we need to understand what a chordal graph is. A Chordal Graph is a graph which does not contain chordless cycles of length greater than three.



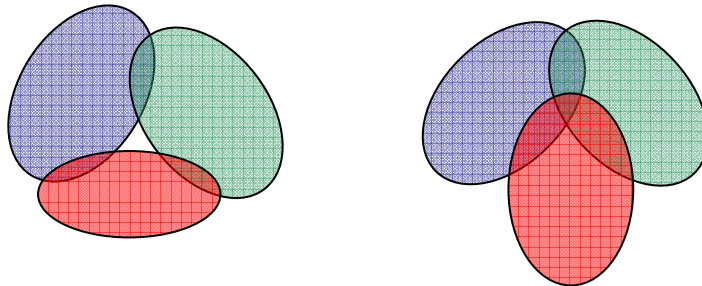
In this figure we see a cycle (black) with two chords (green). As for this part, the graph is chordal. However, removing one green edge would result in a non-chordal graph. Indeed, the other green edge with three black edges would form a cycle of length four with no chords.

Secondly we need to comprehend the concept of Helly property. A set S of sets S_i has the **Helly property** if for every subset T of S the following hold: if the elements of T pairwise intersect, then the intersection of all elements of T is also non-empty.

E.g. A family $\{T_i \mid i \in I\}$ of subsets of a set T is said to satisfy the **Helly property**

if, for any collection of sets from this family, $\{T_i \mid j \in J \subseteq I\}$, $\bigcap_{j \in J} T_j = \emptyset$,

whenever $T_j \cap T_k \neq \emptyset, \forall j, k \in J$.



The picture on the left doesn't satisfy the Helly property but the picture on the right does.

Let us try to intuitively understand this concept. In the figure on the left, the 3 sets (red, blue, green) pairwise intersect. i.e.

$$\text{red} \cap \text{blue} \neq \emptyset$$

$$\text{red} \cap \text{green} \neq \emptyset$$

$$\text{blue} \cap \text{green} \neq \emptyset$$

$$\text{But, red} \cap \text{blue} \cap \text{green} = \emptyset$$

Hence it does not satisfy Helly property.

For graph analysis we check for each of the above 2 cases being true or false and accordingly deduce based on the following 2 theorems:

1. There exists a conservative Dollo parsimony tree for a given set of multidomain architectures, iff the domain overlap graph for this set is chordal
2. There exists a static Dollo parsimony tree for a set of multidomain proteins, iff the domain overlap graph for this set is chordal and satisfies the Helly property

We now move back and try to answer the questions stated in the multidomain protein mystery. It turns out that for small and medium size superfamilies, independent merging of protein domains is rare in contrast to that of large and complex superfamilies. Besides, once created, domain architectures persist in small and medium superfamilies but not in large complex ones.

10. Experimental Results

Table 1. The percentage of superfamilies that are consistent with the perfect phylogeny (PP), static Dollo parsimony (SDP) and conservative Dollo parsimony (CDP) criteria. Abbreviations: PA - preferential attachment; NE - not estimated

set	# super-families	% PP	% SDP	% CDP	% random uniform	% random PA
Mouse	983	95	99	99.7	NE	NE
Mouse.c.4-5	88	99	100	100	80	98
Mouse.c.6-8	37	84	100	100	31	66
Mouse.c.9-10	11	66	100	100	17	25
Mouse.c.11-20	23	31	96	96	1.7	1.0
Mouse.c.21-30	9	0	66	100	0	0
Mouse.c.31- *	8	0	50	75	0	0
Nr90	2896	80	98	99.9	NE	NE
Nr90.c.4-5	143	57	99	99.5	80	98
Nr90.c.6-8	130	37	99	100	31	66
Nr90.c.9-10	40	28	100	100	17	25
Nr90.c.11-20	104	13	87	99	1.7	1.0
Nr90.c.21-30	34	6	53	88	0	0
Nr90.c.30- *	28	0	15	50	0	0
Human Kin	101	11	100	100	NE	NE

PP: one change allowed

SDP: harder to gain easier to loose

CDP: like static dollo parsimony but pairs of characters

\

Smaller families are always explained by dollo but not by perfect phylogeny when we move to larger families. If the families are big, all models are having a hard time to explain the data and require more complex ways to evolve. The columns with random percentage displays the probability of explaining evolution of the multi domain proteins by chance.