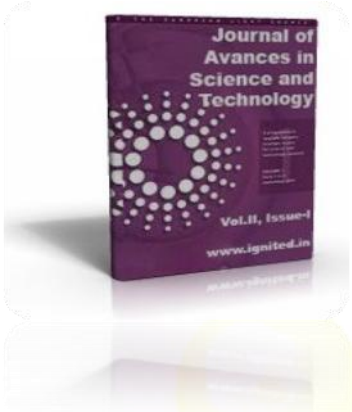
ISSN-2230-9659

Theoretical Study of Graph of Multidomain Proteins



Indu Rathi
Research Scholar, Singhania University,
Rajasthan, India

ABSTRACT:-

We study properties of multidomain proteins from a graph theoretical perspective. In particular, we demonstrate connections between properties of the domain overlap graph and certain variants of Dollo parsimony models. We apply our graph theoretical results to address several interrelated questions: do proteins acquire new domains infrequently, or often enough that the same combinations of domains will be created repeatedly through independent events? Once domain architectures are created, do they persist? In other words, is the existence of ancestral proteins with domain compositions not observed in contemporary proteins unlikely? Our experimental results indicate that independent merges of domain pairs are not uncommon in large super-families.

ISSN-2230-9659

1 INTRODUCTION

Protein domains are elementary units of protein structure and evolution. About two thirds of proteins in prokaryotes and eighty percent in eukaryotes are multidomain proteins [1]. On average, a protein has two to three domains, but there are proteins for which the domain count exceeds one hundred [15,31].

There is no agreement on a precise definition of protein domain. The definition adopted in this work assumes that domains are conserved evolutionary units that are (1) assumed to fold independently, (2) observed in different proteins in the context of different neighboring domains, and are (3) minimal units satisfying (1) and (2).

Multidomain proteins pose a challenge in the analysis of protein families. Traditional approaches for studying the evolution of sequences were not designed with multidomain proteins in mind. For example, gene family evolution is typically modeled as a tree built from multiple sequence alignment. However, it is not clear how to construct such an alignment for a family with heterogeneous domain composition. Another challenge arises in graph theoretical approaches to protein family classification [22,19,34]. This approach typically models the protein universe as a similarity graph, G = (V, E), where V is the set of all amino acid sequences and two vertices are connected by an edge if the associated sequences have significant similarity. The idea is first to identify all pairs of homologous proteins and then apply a clustering technique to construct protein families. In an ideal world, protein families would appear as cliques in such a graph, where every member of the family is related to all other members and to no other protein. However, relationships in this graph are not always transitive. First, it may be impossible to detect sequence homology between related but highly diverged sequences. In addition, lack of transitivity can result from domain chaining in multidomain proteins. A protein containing domain *A* is a neighbor of a protein containing domains *A* and *B*, which in turn is connected to a protein containing only domain *B*, but there would be no direct relationship between the proteins containing only *A* and only *B*, respectively. Consequently, in the presence of multidomain proteins, protein families

ISSN-2230-9659

identified by graph clustering methods may contain completely unrelated proteins. More methods that deal explicitly with multidomain proteins are needed.

In order to focus on the properties of multidomain proteins and the relationships between them, we introduce the *protein overlap graph* and its dual, the *domain overlap graph.* In the protein overlap graph, the vertices are proteins represented by their domain architectures, where domains are represented by probabilistic models of multiple sequence alignments, such as PSSMs [14] or HMMs [5, 24]. Two vertices are connected by an edge if the corresponding proteins share a domain. In the domain overlap graph, the vertices are protein domains and two domains are connected by an edge if there is a protein that contains both domains. These abstractions allow us to focus on domain architectures.

In the current work, we study the structure of domain overlap graphs to gain insight into evolution of multidomain architectures. Multidomain proteins can be formed by gene fusion [20,23,32], domain shuffling [1,4,25,27] and retro- transposition of exons [26]. We abstract these biological mechanisms into two operations: domain merge and domain deletion. We use the term domain merge to refer to any process that unites two or more previously separate domains in a single protein. Domain deletion refers to any process in which a protein loses one or more domains. We represent a domain architecture by the set of its domains. Obviously, this abstraction neglects the fact that multidomain proteins are also subject to domain rearrangement, tandem duplication, and sequence divergence. However in the case of domain pairs it has been observed that only about 2% of such pairs occur in both possible orders [4]. Nevertheless, we must keep in mind our simplifying assumptions while interpreting the results.

We apply the graph theoretic tools developed in this paper to genomic data to consider two questions: First, is domain merging a rare event or is it common for the same pair of domains to arise repeatedly through independent events? Second, once domain architectures are created do they persist? In other words, do the majority of ancestral architectures occur as subsets of some contemporary protein architectures? It has been argued that the vertex degree for domain overlaps

ISSN-2230-9659

graphs can be reasonably approximated by power law [33,2]. The most popular method of modeling such distribution is using the preferential attachment model [3]. Can this model be applied to multidomain proteins? We investigate these questions using the following approach:

1. We define two parsimony models for multidomain family evolution based on the concept of Dollo parsimony, which we call conservative and static Dollo parsimony. The existence of a conservative Dollo parsimony for a protein family is consistent with a history in which every instance of a domain pair observed in contemporary members of the family arose from a single merge event. The existence of a static Dollo parsimony is consistent with a history in which no ancestor contains a domain combination not seen in a contemporary taxon.

2. We establish a relationship between these parsimony models and particular structures in the domain overlap graph, namely chordality and the Helly property. (Rigorous definitions of these concepts are given in the body of the paper.)

3. We adapt fast algorithms for testing chordality and the Helly property previously developed by other authors to obtain fast existence tests for conservative and static Dollo parsimony and reconstruction of corresponding trees.

4. Using a result from random graph theory, we design a method for selecting a statistically informative test set. We also test the agreement of preferential attachment model with the data.

5. We apply these tests to genomic data and determine the percentage of protein families that can be explained by static or conservative Dollo parsimony.

The paper is organized as follows. First, we review the relevant phylogenetic models and introduce our restrictions on the Dollo parsimony in Section 2. In Section 3, we introduce the graph theoretical concepts used in the paper and show how they apply to the domain overlap graph. We also provide an elegant link between these concepts and parsimony models introduced

ISSN-2230-9659

in Section 2. The application of the theoretical results to genomic data is presented in Section 4. Finally, we provide conclusions and directions for future research.

2 EXPERIMENTAL RESULTS

We apply the methods developed in the previous section to genomic data sets to investigate the questions stated in the introduction:

1. Is independent merging of the same pair of domain a rare event?
2. Do domain architectures persist through evolution?

To do this, we divide the protein universe into overlapping sets of proteins called superfamilies. Each domain defines one superfamily, namely the set of all proteins that contain the given domain. For example, all proteins containing the kinase domain form one superfamily, proteins containing the SH2 domain form another superfamily and these two superfamilies intersect. It is important for our argument that each superfamily have a common reference point - here the common domain. This reference point allows us to interpret each merge as an insertion with respect to this domain. In particular, multiple independent insertions correspond to multiple independent merges of the inserted domain and the reference domain. For each superfamily in our data set, we determine whether it satisfies the perfect phylogeny and conservative and static Dollo criteria. To estimate the significance of our results, we also investigate the probability of observing conservative Dollo parsimony in two null models, uniform random graphs (Erdocs-Renyi model)and random scale free graphs generated using preferential attachment random model.

Null Models. The existence of a conservative Dollo parsimony tree for a given domain superfamily is a necessary but not a sufficient condition for concluding that no repeated, independent merges occurred in the history of the family. We therefore estimate the probability that a superfamily admits a conservative Dollo phylogeny by chance under two different null models. Note, that this is equivalent to determining the probability that a graph of with a given number of vertices is chordal under our null hypotheses.

ISSN-2230-9659

All graphs with less than four vertices are chordal, as are all acyclic graphs (i.e., graphs which are collections of trees). Since a random, sufficiently sparse graph will be acyclic with high probability, such a graph is also likely to be chordal. In fact, a random graph with edge probability $p < n$, where n is number of vertices, is almost certainly acyclic when $c < 1$, while almost all vertices of such a graph belong to a cycle when $c > 1$ and the phase transition occurs at $p = n$ [7]. Consequently, since we are interested in graphs that are unlikely to be chordal by chance, we consider only graphs with at least four vertices that have at least as many edges as vertices. We define a *complex superfamily* to be a superfamily whose domains overlap graph satisfies these criteria and restrict our analysis to complex superfamilies in our data sets. To determine the probability of observing conservative Dollo parsimony in complex superfamilies by chance, we collected statistics to estimate the value of c for domain overlap graphs in our data set. We then used simulation (1000 runs) to estimate the probability that a random graph with uniform edge probability $p = \frac{c}{n}$ is chordal.

Several papers have suggested that the domain overlap graphs have scale free properties [2,33]. We therefore also considered a null model based on preferential attachment, a classical random model for scale free graphs [3]. Under this model, a random graph is constructed iteratively. At each step, a new vertex is connected to an existing vertex with probability proportional to the degree of that vertex. We simulated the preferential attachment model taking care that the parameters are chosen in such a way that the edge density of the resulting random graphs is approximately the same as that in domain overlap graphs of the same size.

Data. We use two different data sets derived from SwissProt version 44 released in 09/2004 [6] (http://us.expasy.org/sprot/). The first contains all mouse proteins, thus all homologous proteins in this set are paralogs. In contrast, the second test set consists of all non redundant (nr90) proteins in SwissProt, and thus contains both paralogs and orthologs. The architectures of each protein in both sets were identified using CDART [14] based on PSSM domain models. The domains identified by CDART as similar have been clustered using single linkage clustering and subsequently considered as one *superdomain.* The proteins that contained no recognizable domain

ISSN-2230-9659

were removed, leaving 256,937 proteins with 5,349 distinct domains in the nr90 data set and 6,681 proteins with 1951 distinct domains in the mouse data set. Of these, 2,896 nr90 and 983 mouse superfamilies have at least one partner domain. We let Mouse.c and nr90.c denote the set of complex superfamilies in mouse and nr90, respectively. To determine the effect of superfamily size on the results, we defined Mouse.c.x-y and nr90.c.x-y to be sets of superfamily in Mouse.c and nr90.c, respectively, containing at least *x* and at most *y* domains.

There is always a danger of inaccuracy when working with large, automatically annotated, data sets. Since errors in domain architecture identification could result in incorrect conclusions concerning domain insertion and loss, we also tested our approach on on a hand curated data set, namely the kinase superfamily, which has been heavily studied and for which it is possible to obtain highly reliable domain annotations. We compared the set of complete human protein sequences, obtained from SwissProt along with their symbols and Pfam codes, with a list of designated kinase gene symbols and Pfam codes (PF00069, PF001163 and PF01633) derived from three recent, genomic analyses of the kinase superfamily [30,18,8]. A protein was judged to be a kinase if it was annotated with a known kinase gene symbol or Pfam code. This procedure resulted in a set of 378 human kinase sequences. The domain architectures of these kinases were then obtained from CDART [14]. From this curated set, we analyzed the kinase superfamily, and all superfamilies that overlapped with it.

Analysis. To test the consistency of the data with the perfect phylogeny, static Dollo parsimony, and conservative Dollo parsimony models, we implemented the algorithms discussed in the previous sections using the LEDA platform [29].

ISSN-2230-9659

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

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

The agreement with perfect phylogeny criterion was tested using compatibility criterion [12]. To test conservative Dollo parsimony, we implemented a chordality test and for static Dollo parsimony we additionally tested if the Helly property is satisfied. Using these tools, we test our data for these criteria and asked under what circumstances could at least 90% of superfamilies be explained by a given evolutionary model. The results are summarized in Table 1.

Not surprisingly, with the exception of very small (in terms of number of different domains or equivalently the size of domain overlap graph) superfamilies in mouse perfect phylogeny does not meet this standard suggesting that it is not a suitable model for multidomain protein evolution. In contrast, 95% or more of complex superfamilies up to size 20 in mouse and size 10 in nr90 could be explained by static Dollo parsimony. All but the largest complex superfamilies (greater than 30 in mouse and greater than 20 in nr90) were consistent with conservative Dollo parsimony. In contrast, the probability of observing conservative Dollo parsimony by chance was much lower in both null models. Furthermore, our results show that domain overlap graphs of real multidomain

ISSN-2230-9659

superfamilies do not have the same the topological structure as random scale free graphs of the same size and edge density constructed according to preferential attachment random model.

While the vast majority of small and medium size superfamilies admit conservative and static Dollo parsimony, a significant percentage large superfamilies do not. A less restrictive evolutionary model that allows multiple insertions is needed to explain the data. Furthermore, our simplifying assumptions may result in underestimation of the number of independent merges since only merges that violate chordality are detected. For the mouse data set, the superfamilies that do not satisfy conservative Dollo parsimony are FER2, Trypsin, and EGF. For nr90, this set contains 34 superfamilies including TRK, IG, PH, EGF, SH3, C2, and a large superdomain containing several ATPases (the largest superfamily in the nr90 set). Several of these are known to be "promiscuous" domains, which also supports the hypothesis of repeated independent merges in large families [28]. While the quality of domain recognition and incompleteness of the data may be affecting our results, the results for the curated kinases family are consistent with the results for non-curated data (the sizes of all but one domain overlap graphs for this set, are less than 20).

3. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we formulated two new parsimony models and showed their connection to properties of domain overlap graphs. Previous analysis of these graphs focused on counting vertex degrees and statistical analysis of connectivity [2,33]. We demonstrated that these graphs frequently have interesting topological properties, and in fact the topology of domain overlap graphs can provide information about evolution of a multidomain protein family. We applied our new graph theoretical tools to test whether independent merging of the same pair of domains is a rare event and whether domain architectures persist through evolution? In the case of small and medium sizes superfamilies, the data is consistent with this hypothesis. However, our results do not support the hypothesis in the case of large families. We also demonstrate that the topological properties of domain overlap graphs of multidomain superfamilies are very different from those of random scale free graphs of the same size and density. Based on these results, we reject

ISSN-2230-9659

preferential attachment as a mechanism for multidomain protein evolution. This also prompts the question: what evolutionary model for multidomain proteins will explain the observed behavior?

We show that the independent domain mergers can be detected by testing if the corresponding domain overlap graph is chordal. An intriguing question is whether the minimal set of domains which must be removed to obtain a chordal domain overlap graph is related to the set of does this minimal set tend to be promiscuous domains.

Although the focus of this study is evolution of protein architectures, applicability of the methods developed in this paper goes beyond the analysis of multidomain protein superfamilies. They can be applied to analysis of any set of taxa with binary character states.

Another interesting direction of future research is to study of properties of protein overlap graphs. While the domain overlap graph is dual to the protein overlap graph, this duality is not symmetric. Given a protein overlap graph, we can construct the corresponding domain overlap graph, but given a domain overlap graph we cannot reconstruct the initial protein overlap graph. The domain overlap graph thus contains less information than the protein overlap graph.

Therefore, direct analysis of protein overlap graphs may bring new insights in analyzing evolution of multidomain proteins.

REFERENCES

1) G. Apic, J. Gough, and S.A. Teichmann. Domain combinations in archaeal, eubac- terial and eukaryotic proteomes. *J Mol Biol*, 310:311-325, 2001.

2) G. Apic, W. Huber, and S.A. Teichmann. Multi-domain protein families and domain pairs: Comparison with known structures and a random model of domain recombination. *J. Struc. Func. Genomics*, 4:67-78, 2003.

3) A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509-512, 1999.

4) M. Bashton and C. Chothia. The geometry of domain combination in proteins. *J Mol Biol*, 315:927-939, 2002.

ISSN-2230-9659

5) Bateman, E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, and E.L. Sonnhammer. The Pfam protein families database. *Nucleic Acids Res.*, 28(1):263-266, 2000.

6) Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.,* 31:365-370, 2003.

7) Bollobas. *Random Graph Theory.* Cambridge University Press, 2001.

8) S. Cheek, H. Zhang, and N. V. Grishin. Sequence and structure classification of kinases. *J Mol Biol,* 320(4):855-881, Jul 2002.

9) L. Danzer, B. Grunbaum, and V. Klee. Helly's theorem and its relatives. *Convexity, AMS,* 7:101-180, 1963.

10) W.H.E. Day, D. Johnson, and D. Sankoff. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences,* 81:33-42, 1986.

11) D.Gusfield. Efficient methods for inferring evolutionary history. *Networks,* 21:1928, 1991.

12) J. Felsenstein. *Inferring Phylogenies.* Sinauer Associates, 2004.

13) F. Gavril. The intersection graphs of subtrees in trees are exactly the chordal graphs. *J. Comb. Theory (B),* 16:47-56, 1974.

14) L.Y. Geer, M. Domrachev, D.J. Lipman, and S.H. Bryant. CDART: protein homology by domain architecture. *Genome Res.,* 12(10):1619-23, 2002.

15) M. Gerstein. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold des.*, 3:497-512, 1998.

16) M. Golumbic. *Algorithmic Graph Theory and Perfect Graphs.* Academic Press, New York, 1980.

17) J. Gu and X. Gu. Natural history and functional divergence of protein tyrosine kinases. *Gene,* 317:49-57, 2003.

18) S.K. Hanks. Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. *Genome Biol,* 4(5):111, 2003.

19) Heger and L. Holm. Exhaustive enumeration of protein domain families. *J. Mol Biol,* 328:749-767, 2003.

ISSN-2230-9659

20) I.Yanai, Y.I. Wolf, and E.V. Koonin. Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol*, 3, 2002. research:0024.

21) J.S.Farris. Phylogenetic analysis under Dollo's law. *Systematic Zoology*, 26(1):77- 88, 1977.

22) Krause, J. Stoye, and M. Vingron. The SYSTERS protein sequence cluster set. *Nucleic Acids Res.*, 28(1):270-272, 2000.

23) S. Kummerfeld, C. Vogel, M. Madera, and S. Teichmann. Evolution of multi- domain proteins by gene fusion and fission. *ISMB 2004*, 2004.

24) Letunic, L. Goodstadt, N.J. Dickens, T. Doerks, J. Schultz, R. Mott, F. Ciccarelli, R.R. Copley, C.P. Ponting, and P. Bork P. Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, 31(1):242-244, 2002.

25) Y. Liu, M. Gerstein M, and D.M. Engelman. Evolutionary use of domain recombination: a distinction between membrane and soluble proteins. *Proc Natl Acad Sci USA*, pages 3495 - 3497, 2004.

26) M. Long. Evolution of novel genes. *Curr Opin Genet Dev*, 11(6):673-680, 2001.

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

28) E.M. Marcotte, M. Pellegrini, H.L. Ng, D.W. Rice, T.O. Yeates, and D. Eisen- berg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285:751-53, 1999.

29) K. Mehlhorn and S. Naher. *The LEDA Platform of Combinatorial and Geometric Computing.* Cambridge University Press, 1999.

30) D.R. Robinson, Y.M. Wu, and S.F. Lin. The protein tyrosine kinase family of the human genome. *Oncogene*, 19(49):5548-5558, 2000.

31) S.A.Teichmann, J. Park, and C. Chothia. Structural assignments to the my- coplasma genitalium proteins show extensive gene duplications and domain rearrangements, 1998.

32) B. Snel, P. Bork, and M. Huynen. Genome evolution gene fusion versus gene fission. *Trends Genet*, 16:9-11, 2002.

33) S. Wuchty. Scale-free behavior in protein domain networks. *Mol. Biol. Evol.*, 18:1694-1702, 2001.

ISSN-2230-9659

34) G. Yona, N. Linial, and Linial M. Protomap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins: Structure, Function and Genetics*, 37:360-378, 1999.