Published in final edited form as:
J Comput Biol. 2006 March ; 13(2): 351–363.

Graph Theoretical Insights into Dollo Parsimony and Evolution of Multidomain Proteins

Teresa Przytycka*
National Center for Biotechnology Information, US National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA. przytyck@mail.nih.gov

George Davis
Program in Computation, Organizations and Society, Carnegie Mellon University, Pittsburgh, PA, 15213, USA. gbd@andrew.cmu.edu

Nan Song
Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, 15213, USA. nsong@andrew.cmu.edu

Dannie Durand
Departments of Biological Sciences and Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. durand@cmu.edu

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 superfamilies.

Keywords

multidomain proteins; protein evolution; Dollo parsimony; domain overlap graph; chordal graphs; Helly property

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 (Apic et al., 2001). On average, a protein has two to three domains, but there are proteins for which the domain count exceeds one hundred (Gerstein, 1998; Teichmann et al., 1998).

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).

*phone: (301) 496-1723, fax: (301) 480-4637

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 (Krause et al., 2000; Heger and Holm, 2003; Yona et al., 1999). 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 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 (Geer et al., 2002) or HMMs (Bateman et al., 2000; Letunic et al., 2002). 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 (Yanai et al., 2002; Kummerfeld et al., 2004; Snel et al., 2002), domain shuffling (Apic et al., 2001; Bashton and Chothia, 2002; Liu et al., 2004; Patthy, 1999) and retrotransposition of exons (Long, 2001). 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 (Bashton and Chothia, 2002). 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 graphs can be reasonably approximated by power law (Wuchty, 2001; Apic et al., 2003). Graphs with this property (often called “scale free”) are commonly modeled using a

preferential attachment model (Barabasi and Albert, 1999). We use the following approach to determine whether this technique can be applied to the study of multidomain proteins:

1. We define two parsimony models for multidomain family evolution based on the concept of Dollo parsimony, which we call Conservative Dollo and Static Dollo parsimony respectively. 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 existing fast algorithms for testing chordality and the Helly property to obtain fast existence tests for conservative and Static Dollo phylogeny and reconstruction of corresponding trees. We note that intersection tree representation provides a sample Conservative Dollo tree and prove that finding a Conservative Dollo tree that minimizes the number of deletions is NP-complete.
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 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 Tree models

Gene family evolution is traditionally modeled by phylogenetic trees, where leaves are sequences and internal nodes represent either speciation or duplication events. Gene trees are traditionally built from multiple sequence alignments (MSAs). However, it is not clear how to construct an MSA for a family with heterogeneous domain composition. One approach is to use the MSA of one domain only (see, for example, (Gu and Gu, 2003; Robinson et al., 2000)). There is no guarantee, however, that the resulting tree will capture large scale changes in domain composition. Therefore, in this work we will consider parsimony models, where the primary evolutionary events are domain insertion and deletions.

In general, parsimony methods assume that each taxon is characterized by a set of characters or attributes. Each character can assume a finite set of possible states and can change state in the course of evolution. The maximum parsimony tree is a tree with leaves labeled with character states associated with the input taxa, and internal nodes labeled with the inferred character states of ancestral taxa, such that the total number of character changes along its branches is minimized. Additional restrictions on the type, number and direction of changes lead to a variety of specific parsimony models (Felsenstein, 2004). In this work, we focus on binary characters, characters that take only the values zero or one, usually interpreted as the presence or absence of the attribute in the taxa.

The most restrictive parsimony assumption is perfect phylogeny: a tree in which each character state change occurs at most once (Felsenstein, 2004). One method of testing the existence of perfect phylogeny tree is based on the compatibility criterion. For a given set of taxa, two characters *A* and *B* are compatible if and only if there do not exist four different taxa respectively representing all four possible combination of character states for *A* and *B* (that is, (0, 0), (0, 1), (1, 0), (1, 1)). The appearance of all four combinations indicates that one of the two characters must have changed state twice. A set of taxa admits a perfect phylogeny if and only if every pair of taxon is compatible (Felsenstein, 2004).

In Dollo parsimony, a character may change state from zero to one only once, but from one to zero multiple times (Farris, 1977). This model is appropriate for complex characters such as restriction sites and introns which are hard to gain but relatively easy to lose (Felsenstein, 2004). In Camin-Sokal parsimony, no transition from derived state to ancestral state is allowed (Camin and Sokal, 1965). That is if ancestral state is $O = (0,... 0)$ then no character change back to 0 is allowed.

We model multidomain protein evolution in terms of domain insertion and loss. In our model, the taxa are domain architectures and each domain defines a single binary character, where state one corresponds to the presence, and zero to the absence, of the domain in a given architecture. Thus, a state change from zero to one corresponds to an insertion and from one to zero to a deletion. This model focuses on the evolution of domain architecture, ignoring sequence evolution and thus obviating the problem of constructing an appropriate MSA for tree reconstruction. Figure 1 (a) shows a domain architecture phylogeny for the protein tyrosine kinases, based on a tree constructed from an MSA of the kinase domain (Robinson et al., 2000). Note 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 lingand binding domains) the number of insertions and deletions would be smaller.

The general maximum parsimony and the Dollo parsimony problems are optimization problems: an optimal tree satisfying the given parsimony criterion is sought. In contrast, the perfect phylogeny problem asks whether a perfect phylogeny exists. If such a tree does exist, it is guaranteed to be optimal. Finding the most parsimonious tree (both in the general setting as well with the Dollo restriction) is NP-complete (Day et al., 1986). The existence of a perfect phylogeny can be solved in $O(nm)$ time, where m is the number of characters (i.e., domains) and n is the number of taxa (i.e., architectures) (Gusfield, 1991).

In contrast to perfect phylogeny, it is always possible to construct a Dollo phylogeny by positing an ancestral taxon where the state of every character is one. Since there is no restriction on the number of transitions from one to zero, any combination of character states found in the leaf taxa can be generated. Such a tree makes no sense in the context of multidomain evolution, since it implies the existence of an ancient protein containing all domains seen in any contemporary protein in the family. Can we put a restriction on the Dollo phylogeny so that the existence of such restricted Dollo phylogeny is both informative and computationally tractable? In this paper we propose two such restrictions:

Static Dollo Parsimony is a Dollo parsimony with the following restriction: for any ancestral taxon the set of characters in state one in this taxon is a subset of the set of characters in state one in some leaf taxon (hence, the term “static”). We assume here that more than one character can change in one step.

Conservative Dollo Parsimony is a Dollo parsimony with the following restriction: for any ancestral taxon and any pair of characters that appear in state one in this taxon, there exists a leaf taxon where these two characters are also in state one.

Clearly every static phylogeny is also conservative. From the perspective of multidomain proteins, the intuition motivating the conservative restriction is as follows. The simultaneous presence of two domains in one protein often suggests that these domains contribute to the functionality of the protein as a pair. For instance, the SH2 and SH3 domains frequently appear together in various signal transduction proteins involving recognition of phosphorylated tyrosine. SH2 domains localize tyrosine phosphorylated sites while SH3 domain binds to target proteins through sequences containing proline and hydrophobic amino acids. If the domains acting in concert offer a selective advantage, it is unlikely that the pair, once formed, would later separate in all contemporary protein architectures. Accordingly, Conservative Dollo parsimony provides a correct parsimony model when the possibility of complimentary domains being separated is excluded. (Note that it may be possible that a pair of domains does not form a functional unit without additional domains but we do not explore such intricate relationships here due to insufficient data). Static Dollo parsimony additionally requires that the set of characters in state “one” in an ancestral taxa is a subset of the set of characters in state one in at least one contemporary taxa. Consequently, an ancestral architecture (defined as a set of domains) is a subset of at least one contemporary architecture.

Unlike the general Dollo parsimony which can be always inferred (even in the case where it is not a reasonable model), a set of taxa may not admit the Conservative Dollo parsimony. Such a failure can be interpreted in two ways: the single insertion assumption is not reasonable, or conservative assumption is too strong. Thus non-existence of Conservative Dollo phylogeny 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.

In this paper, we show that there is an elegant link between existence of Static and Conservative Dollo phylogenies and some graph theoretical properties of the domain overlap graph. This leads to fast algorithms for testing existence of such restricted phylogenies. We also show that computing optimal (in terms of the number of deletions) Conservative and Static Dollo phylogeny is NP-complete.

3 Graph theoretical properties of domain overlap graphs and their relation to restricted Dollo parsimony

In this section we present our theoretical results. We start with the analysis of the domain overlap graph. Stated formally, the domain overlap graph for a given family of multidomain proteins is the graph $G=(V, E)$ such that V is the set of all domains represented in the data base and $(u, v) \in E$ if there exists a protein p in the set such that both u and v appear in p. Below, we state the definition of chordality and discuss its importance in the context of the domain overlap graph. Subsequently, we review the Helly property and its relation to chordal graphs. Finally, we show how these concepts can be exploited to answer the questions stated in the introduction and discuss related statistical issues.

3.1 Chordal graphs and their properties

Chordal graphs constitute an important and well studied graph family (Golumbic, 1980). A chord in a graph is any edge that connects two non-consecutive vertices of a cycle. A chordal graph is a graph which does not contain chordless cycles of length greater than three. Intuitively, any cycle of length greater than three in a chordal graph is triangulated, that it is partitioned (not necessarily uniquely) into triangles using edges of the graph. This motivates another term for chordal graphs, namely triangulated graphs. Figure 2 (a) shows the domain overlap graph for a set of domains in protein kinase family shown in Figure 1. For simplicity, only domains that occur in more than one architecture are used. Note that this graph is chordal.

The important property that is explored directly in this paper is the relation between chordal graphs and trees. To elucidate this relation we need to introduce the concept of intersection graph.

Let F be a family of objects, such as intervals on a coordinate line or rectangles in space. A graph $G = (V, E)$ is called an intersection graph of F if each vertex, $v \in V$, corresponds to an object in F and $(u, v) \in E$ if and only if the objects in F corresponding to u and v intersect.

We will consider a special family of intersection graphs where the objects are subtrees of some (usually unrooted) tree. We will refer to the tree as the guide tree. Here, by a subtree of a tree we understand any connected subgraph of a tree. Furthermore, our family typically does not contain all possible subtrees of a tree. **Theorem**(**Gavrill** (Gavril, 1974)) A graph G is chordal if and only if there exists a tree T and a family of subtrees of this tree such that G is the intersection graph of this family.

Our key observation is stated in the following theorem:

Theorem 1. There exists a Conservative Dollo phylogeny for a given set of multidomain architectures, if and only if the domain overlap graph for this set is chordal.

Proof. ⇒Assume that conservative Dollo phylogeny exists. We take this tree as the guide tree for an intersection graph. The family of intersection subtrees is defined as follows. For any domain consider all nodes in the Dollo tree (leaves or ancestral) that contain this domain. By the Dollo property, these nodes form a connected subtree. Consider the family of such subtrees, one tree per each domain. We argue that the intersection graph of this family of subtrees is exactly the domain overlap graph. By the definition of intersection graph, the nodes of this graph correspond to protein domains and there is an edge between two such domains if and only if there exists a node in the Dollo tree containing both domains. Thus if two domains belong to the same protein they are connected by an edge in the intersection graph. We need to show that if two domains do not occur together in at least one protein architecture, then there is no edge between them in the intersection graph. Assume towards a contradiction that there exists an edge between two domains that do not belong to the same architecture. This means that the corresponding subtrees intersect in an internal node but they don't intersect in a leaf. This contradicts the assumption that the tree is conservative. Thus the intersection graph is exactly equal to the domain overlap graph. By Gavril's theorem, the domain overlap graph is chordal.

⇐Assume that domain overlap graph G is chordal. We use another theorem by Gavril (Gavril, 1974; Buneman, 1974), which states that if a graph is chordal then there exists the so called clique tree. Namely consider the set of maximal cliques of G. Then, under assumption that G is chordal, there exists a tree T where vertices of T are maximal cliques and such that for any vertex v of G all cliques that contain v form a connected subtree. Take tree T as the guide tree. By the definition of clique tree, every vertex v of G defines a subtree of T. Note that the intersection graph of these subtrees is exactly the graph G. From T we obtain a Conservative Dollo tree as follows. Note that the domains that included in any particular domain architecture, D, form a clique in the domain overlap graph. This clique is a subset of some maximal clique. Let D be a maximal clique containing the clique D. If more than one such maximal clique exists, we choose one arbitrarily. By the definition of a clique tree, there exists a node in T that corresponds to D. Let us call D parent node for D and D a child node of D. Obviously, D may be a parent node for several architectures. Now for all children of a given node create an arbitrary tree with the children in the leaves and make it adjacent to D. Note that independently of how such subtrees may be constructed, they contain only domains that are already in a clique (that is, each domain

pair appears together in some architecture) and no new pairs are formed. Thus the resulting tree satisfies the Conservative Dollo parsimony restriction.

Note that Theorem 1 guarantees a fast test for existence of a Conservative Dollo phylogeny. However as illustrated on Figure 2 the tree does not need to be unique. Figure 2 shows a domain overlap graph and a corresponding Static Dollo phylogeny. Note that the order of internal nodes in the SH2/SH3 subtree can be switched without inducing a change in the overlap graph. The proof of theorem 1 presented above is constructive in the sense that it constructs a tree that satisfies the Conservative Dollo restrictions. However, the construction does not ensure that this tree minimizes the number of deletions. As we show below, constructing such an optimal tree remains NP-complete.

Theorem 2. The problem of computing optimal Conservative Dollo parsimony tree is NP-complete.

Our proof on this theorem is based in NP-completeness of Camin-Sokal parsimony (Day et al., 1986). We delay the proof of this theorem until the next subsection, after introducing the Helly property. Below, we state the NP-completeness result we will use in the proof.

Theorem [NP-completeness of Comin-Sokal parsimony](Day et al., 1986). The problem of computing optimal Camin-Sokal tree rooted and ancestral state O =(0,... 0) is NP-complete.

3.2 The Helly property

As mentioned in the previous subsection, it is not always possible to construct a Dollo phylogeny without introducing ancestral nodes with domain compositions not observed in the leaves. For example, if we have three domains *A, B,* and *C* and three proteins *AB, BC, CA* then we cannot construct a Dollo phylogeny without introducing an ancestral protein *ABC*. This property is equivalent to the Helly property, named after the Austrian mathematician Eduard Helly (Danzer et al., 1963):

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, $\{T_j \mid j \in J \subseteq I\}$, $\cap_{j\in J} T_j = \varnothing$, whenever $T_j \cap T_k = \varnothing$, $\forall j, k \in J$.

In Figure 3 (a) shows an example of a family of three sets that do not satisfy the Helly property: Each pair intersects but there is no intersection point common to all three. In contrast, Figure 3 (b) shows an example where the three subtrees A, B, C (respectively with vertices {1,3,4,5}, {2,3,4,6}, and {1,2,3}) of tree *T*. The subtrees pairwise intersect and also have a common intersection point in vertex 3. Thus they satisfy the Helly property. The last fact is true for any set of subtrees of a tree: there is no way to have such subtrees pairwise intersect but not intersect in a common point.

The following theorem is well known and attributed to Helly (Danzer et al., 1963):

Theorem: Helly property for family of subtrees of a tree. If *F* is a family of a subtrees of a tree then any subset of *F* satisfies Helly property.

Consistent with the above definition of the Helly property, we introduce the Helly property for a domain overlap as follows:

Defintion (Helly property for domain overlap graph) A domain overlap graph satisfies the Helly property if and only if for every clique in this graph there exists a protein architecture that contains all domains of this clique.

To see why this definition is consistent with the set theoretical definition for each domain *i* consider set T_i of architectures that contain this domain. Then an edge between *i* and *j* corresponds to $T_i \cap T_j = \varnothing$. Subsequently a clique *J* has property that

all $i, j \in JT_i \cap T_j = \varnothing$. The existence of an architecture that contains all domains in the clique J ensures that $\cap_{j \in J} T_j = \varnothing$ since this architecture belongs to all sets T_j in this clique.

The relation of the Helly property to the Static Dollo parsimony is provided by the following theorem:

Theorem 3. There exists a Static Dollo phylogeny for a set of multidomain proteins F, if and only if the domain overlap graph for this set is chordal and satisfies the Helly property.

Proof. $\Rightarrow$ Assume that there exists a Static Dollo phylogeny. Since static parsimony condition is stronger than conservative parsimony, this implies that the domain overlap graph is chordal. Consider such Conservative Dollo tree for F. Since in Dollo phylogeny, for any domain, the tree nodes that contain this domain form a connected subtree, by the Helly property for family of subtrees of a tree, the subtrees corresponding to nodes of any clique have to intersect in a common point. Therefore in any Dollo tree there must exist a node, leaf or ancestral, that contains all domains in the clique. By the static property, this implies that there must exist a protein architecture in F which contains all these domains. Thus the Helly property for the domain overlap graph follows.

$\Leftarrow$ Assume that the domain overlap graph is chordal and satisfies Helly property. Then exactly same construction as in only if part of the proof of Theorem 1 produces a Static Dollo tree.

Finally, we are ready to prove Theorem 2. In fact we will prove a stronger theorem:

Theorem 2a The problem of testing whether there exists optimal Static Dollo phylogeny that uses at most k attribute changes is NP-complete.

Proof. The problem is clearly in NP. To show that it is NP-complete we use a reduction form Camin-Sokal parsimony for tree rooted at O =(1,..., 1). (This is a symmetric case to O =(0,..., 0) thus it is also NP-complete. For any instance of Camin-Sokal parsimony the corresponding Static Dollo parsimony problem is constructed as follows. Let F the instance of the Camin-Sokal parsimony problem where each element of F is a binary vector of attributes states. Then by Camin-Sokal criterion once the attribute is switched to 0 it cannot be converted to 1 again. Under assumption that root is O =(1,..., 1), the Camin-Sokal phylogeny is a Dollo phylogeny for set $F \cup O$. Since O =(1,..., 1) contains all domains (the domain overlap graph is a clique) this Dollo phylogeny is static. Consequently the problem of finding Camin-Socal phylogeny with root O requiring at most k attribute changes is equivalent to finding Static Dollo phylogeny for $F \cup O$ with at most k attribute changes.

4 Experimental results

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

- Is independent merging of the same pair of domain a rare event?
- 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 has 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 several null models, described below.

Null Models.

The existence of a Conservative Dollo phylogeny 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 several different null models. Note, that this is equivalent to determining the probability that a graph with a given number of vertices is chordal under our null hypotheses.

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 $p < \frac{c}{n}$, 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 $p = \frac{1}{n}$ (Bollobas, 2001). 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 (Apic et al., 2003; Wuchty, 2001). We therefore also considered a null model based on preferential attachment, a classical random model for scale free graphs (Barabasi and Albert, 1999). 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.

As an additional control, we randomized each domain overlap graph by swapping endpoints between randomly selected pairs of edges. Swaps leading to self-loops and multi-edges were rejected. Note that this randomization procedure preserves the degree of vertices in each graph. While it is interesting to see what effect such randomization has on the properties of domain overlap graph, one should keep in mind a drawback of this approach. Namely, a domain that is present in several superfamilies will occur in several domain overlap graphs and typically will have different degree in each of them. Therefore, it will be modeled differently in each superfamily. This differentiation by superfamily may be overly specific in conjunction with our null model, which should be universal for all superfamilies.

Data.

We use two different data sets derived from SwissProt version 44 released in 09/2004 (Boeckmann et al., 2003) (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 (Geer et al., 2002) 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 lists of proteins with assigned superdomians is posted at http://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/multidomain.htm. The proteins that contained no recognizable domain 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 those, only the 2,896 nr90 and 983 mouse superfamilies with at least one partner domain were included in our base datasets. We created additional "complex" datasets restricted to domains in superfamilies matching the complexity criteria described above. To determine the effect of superfamily size on the results, we further separated the databases into several subsets admitting superfamilies according to the number of distinct domains in their domain overlap graph.

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 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 (Robinson et al., 2000; Hanks, 2003; Cheek et al., 2002). 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(Geer et al., 2002). 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 (Mehlhorn and Naher, 1999). The agreement with perfect phylogeny criterion was tested using compatibility criterion (Felsenstein, 2004). 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 with up to 20 distinct domains in their overlap graph in mouse and 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 superfamilies do not have the same 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 TyrKc, IG, PH, EGF, SH3, C2, and a large superdomain containing several ATPases (the largest superfamily in the nr90 set). (The full list is given at http://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/multidomain.htm.) Several of these are known to be “promiscuous” domains, which also supports the hypothesis of repeated independent merges in large families (Marcotte et al., 1999). 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).

5 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 (Apic et al., 2003; Wuchty, 2001). 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 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. If so, what is the relation? 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 the 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.

Acknowledgments

We thank L. Y. Geer and S. H. Bryant (NCBI) for providing the complete set of CDART domain architectures, D. Ullman, R. Jothi, and E. Zotenko for valuable discussions. T.P. was supported by the intramural research program of the National Institutes of Health. D.D., G.D. and N.S. were supported by NIH grant 1 K22 HG 02451-01 and a David and Lucille Packard Foundation fellowship.

References

Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J. Mol. Biol 2001;310:311–325. [PubMed: 11428892]

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

Barabasi A-L, Albert R. Emergence of scaling in random networks. Science 1999;286:509–512. [PubMed: 10521342]

Bashton M, Chothia C. The geometry of domain combination in proteins. J. Mol. Biol 2002;315:927–939. [PubMed: 11812158]

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2000;28:263–266. [PubMed: 10592242]

Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370. [PubMed: 12520024]

Bollobas, B. Random Graph Theory. Cambridge University Press; 2001.

Buneman P. A characterisation of rigid circuit graphs. Discrete Math 1974;9:205–212.

Camin JH, Sokal RR. A method for deducting branching sequences in phylogeny. Evolution 1965;19:311–326.

Cheek S, Zhang H, Grishin NV. Sequence and structure classification of kinases. J. Mol. Biol 2002;320:855–881. [PubMed: 12095261]

Danzer L, Grunbaum B, Klee V. Helly's theorem and its relatives. Convexity, AMS 1963;7:101–180.

Day W, Johnson D, Sankoff D. The computational complexity of inferring rooted phylogenies by parsimony. Mathematical Biosciences 1986;81:33–42.

Farris JS. Phylogenetic analysis under Dollo's law. Systematic Zoology 1977;26:77–88.

Felsenstein, J. Inferring Phylogenies. Sinauer Associates; 2004.

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

Geer LY, Domrachev M, Lipman DJ, Bryant SH. CDART: protein homology by domain architecture. Genome Res 2002;12:1619–23. [PubMed: 12368255]

Gerstein M. How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. Fold des 1998;3:497–512. [PubMed: 9889159]

Golumbic, M. Algorithmic Graph Theory and Perfect Graphs. Academic Press; New York: 1980.

Gu J, Gu X. Natural history and functional divergence of protein tyrosine kinases. Gene 2003;317:49–57. [PubMed: 14604791]

Gusfield D. Efficient methods for inferring evolutionary history. Networks 1991;21:19–28.

Hanks SK. Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. Genome Biol 2003;4:111. [PubMed: 12734000]

Heger A, Holm L. Exhaustive enumeration of protein domain families. J. Mol. Biol 2003;328:749–767. [PubMed: 12706730]

Krause A, Stoye J, Vingron M. The SYSTERS protein sequence cluster set. Nucleic Acids Res 2000;28:270–272. [PubMed: 10592244]

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

Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P. Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res 2002;31:242–244. [PubMed: 11752305]

Liu Y, Gerstein M, Engelman DM. Evolutionary use of domain recombination: a distinction between membrane and soluble proteins. Proc. Natl. Acad. Sci. USA 2004;3495:3497.

Long M. Evolution of novel genes. Curr. Opin. Genet. Dev 2001;11:673–680. [PubMed: 11682312]

Marcotte FM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science 1999;285:751–753. [PubMed: 10427000]

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

Patthy L. Genome evolution and the evolution of exon-shuffling-a review. Gene 1999;238:103–114. [PubMed: 10570989]

Robinson DR, Wu YM, Lin SF. The protein tyrosine kinase family of the human genome. Oncogene 2000;19:5548–5558. [PubMed: 11114734]

Snel B, Bork P, Huynen M. Genome evolution gene fusion versus gene fission. Trends Genet 2002;16:9–11. [PubMed: 10637623]

Teichmann SA, Park J, Chothia C. Structural assignments to the mycoplasma genitalium proteins show extensive gene duplications and domain rearrangements. Proc. Natl. Acad. Sci. USA 1998;95:14658–14663. [PubMed: 9843945]

Wuchty S. Scale-free behavior in protein domain networks. Mol. Biol. Evol 2001;18:1694–1702. [PubMed: 11504849]

Yanai I, Wolf YI, Koonin EV. Evolution of gene fusions: horizontal transfer versus independent events. Genome Biol 2002;3Research:0024

Yona G, Linial N, 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 1999;37:360–378.

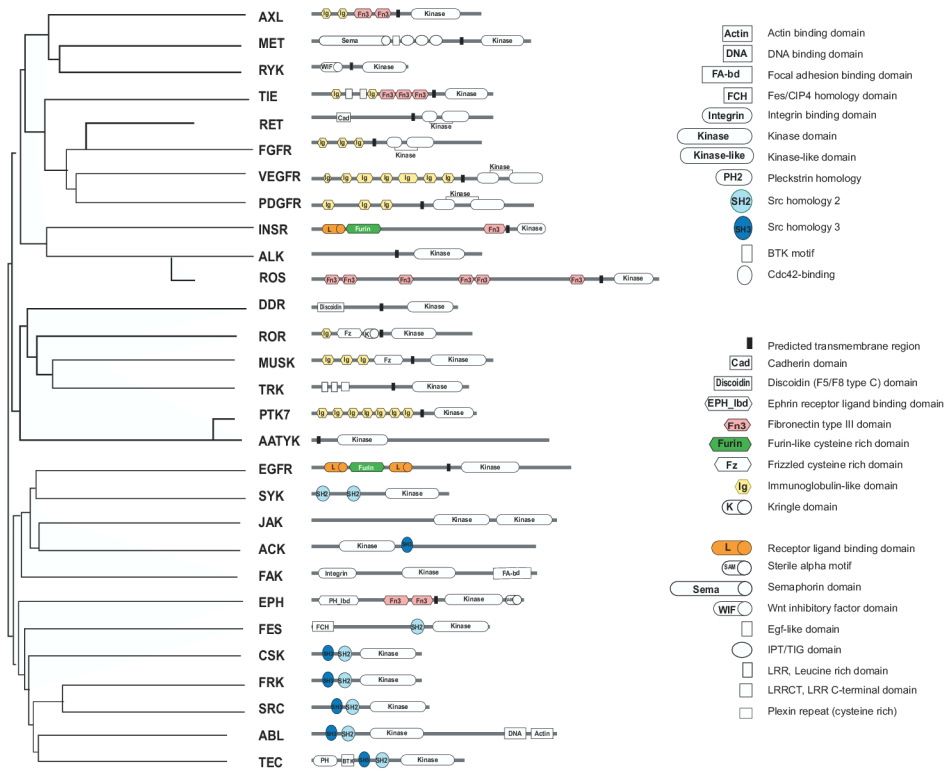

Figure 1.
Phylogenetic tree of family protein tyrosine kinase family, adopted from the tree presented in (Robinson et al., 2000) constructed from an MSA of the kinase domain.

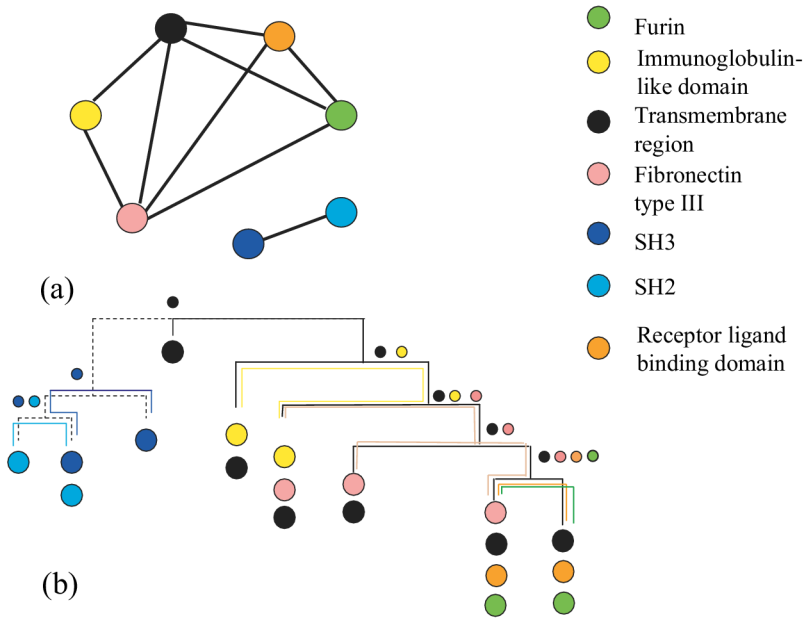

Figure 2.

a) The domain overlap graph for selected domains from the human tyrosine kinase family. Chosen domains belong to more than one architecture (under assumption that architectures containing the same set of domains are considered to be the same). The kinase domain is omitted since it is present in all these architectures. b) Representation of the domain overlap graph as an intersection graph of subtrees of a tree. The correspondence between subtrees and domains is indicated by corresponding colors. The label of a node indicates which subtrees intersect in this node.

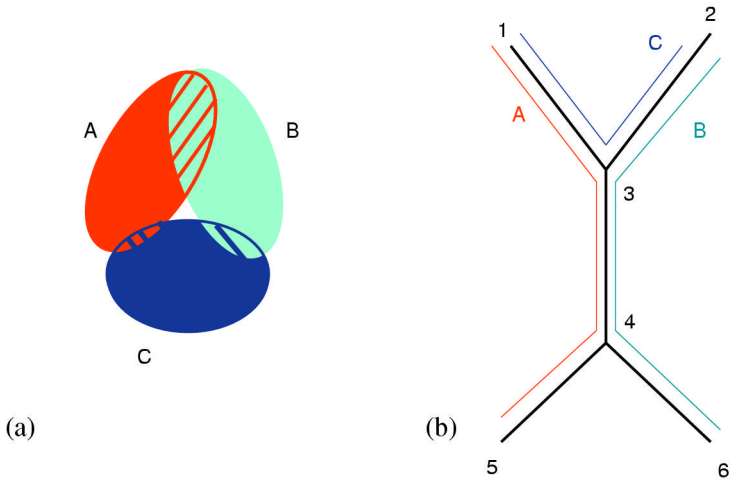

Figure 3.
(a) Three ovals that do not satisfy the Helly property; (b) and three subtrees of a tree which satisfy the Helly property.

Table 1

The percentage of superfamilies that are consistent with the perfect phylogeny (PP), Static Dollo parsimony (SDP) and conservative Dollo parsimony (CDP) criteria. Databases marked ”(c)” include only super-families meeting complex criteria. ”num domains” indicates that only superfamilies with the specified number of domains in their domain overlap graph were included. ”num super” indicates the number of superfamilies meeting these criteria. % Unif, % PA, and %FD refer to the expected percentage of families admitting a Conservative Dollo parsimony under the uniform, preferential attachment, and fixed degree (endpoint swapping) null models. NE indicates a quantity not estimated.

Database	num domains	num super	%PP	%SDP	%CDP	%Unif	%PA	%FD
Mouse	*	983	95	99	99.7	NE	NE	NE
Mouse (c)	4-5	88	99	100	100	80	98	100
Mouse (c)	6-8	37	84	100	100	31	66	73
Mouse (c)	9-10	11	66	100	100	17	25	54
Mouse (c)	11-20	23	31	96	96	1.7	1.0	21
Mouse (c)	21-30	9	0	66	100	0	0	0
Mouse (c)	31-*	8	0	50	75	0	0	0
Nr90	*	2896	80	98	99.9	NE	NE	NE
Nr90 (c)	4-5	143	57	99	99.5	80	98	99
Nr90 (c)	6-8	130	37	99	100	31	66	89
Nr90 (c)	9-10	40	28	100	100	17	25	74
Nr90 (c)	11-20	104	13	87	99	1.7	1.0	26
Nr90 (c)	21-30	34	6	53	88	0	0	2.9
Nr90 (c)	30-*	28	0	15	50	0	0	0
Human Kin	*	101	11	100	100	NE	NE	NE