An Important Connection Between Network Motifs and Parsimony Models

Teresa M. Przytycka

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

Abstract. We demonstrate an important connection between network motifs in certain biological networks and validity of evolutionary trees constructed using parsimony methods. Parsimony methods assume that taxa are described by a set of characters and infer phylogenetic trees by minimizing number of character changes required to explain observed character states. From the perspective of applicability of parsimony methods, it is important to assess whether the characters used to infer phylogeny are likely to provide a correct tree. We introduce a graph theoretical characterization that helps to select correct characters. Given a set of characters and a set of taxa, we construct a network called character overlap graph. We show that the character overlap graph for characters that are appropriate to use in parsimony methods is characterized by significant under-representation of subnetworks known as holes, and provide a mathematical validation for this observation. This characterization explains success in constructing evolutionary trees using parsimony method for some characters (e.g. protein domains) and lack of such success for other characters (e.g. introns). In the latter case, the understanding of mathematical obstacles to applying parsimony methods in a direct way has lead us to a new approach for dealing with inconsistent and/or noisy data. Namely, we introduce the concept of persistent characters which is similar but less restrictive than the well known concept of pairwise compatible characters. Application of this approach to introns produces the evolutionary tree consistent with the Coelomata hypothesis. In contrast, the direct application of a parsimony method, using introns as characters, produces a tree which is inconsistent with any of the two competing evolutionary hypotheses. Similarly, replacing persistence with pairwise compatibility does not lead to a correct tree. This indicates that the concept of persistence provides an important addition to the parsimony metohds.

1 Introduction

The term *biological network* is used in connection to any network where nodes correspond to biological entities (like proteins, genes, metabolites, etc.) and edges are defined by a particular relation between these biological units. Can such biological networks help us to understand evolutionary processes? A number of

A. Apostolico et al. (Eds.): RECOMB 2006, LNBI 3909, pp. 321–335, 2006.

studies have focused on the scale free property – a characteristic power-law like distribution of node degrees observed in various biological networks [4]. However, it has been demonstrated [26, 21] that different evolutionary mechanisms can lead to non-distinguishable scale free-like characteristics. Thus, analysis of degree distribution alone does not bring sufficient insight into the evolution of a network. Recently, small size subgraphs, termed *network motifs*, attracted significant attention [23,34,22,24]. The idea is to consider, exhaustively, all possible subnetworks up to a certain size and identify network motifs which are present more frequently than expected by chance.

In this work, we introduce the concept of a character overlap graph and relate the frequency of occurrences of certain network motifs in these graphs to the evolution of the corresponding character traits. Consider a set of taxa, where each taxon is described by a vector of attributes, the so called *characters*. Assume that each character can assume binary values: one – if the taxon has the property described by the character (we will simply say that the taxon contains the character) and zero – otherwise. We further assume that during the evolution characters are gained and/or lost. This acquisition and loss of character traits is the basis for inferring evolutionary trees using parsimony methods. Maximum parsimony methods search for the evolutionary tree with the topology that can explain the observed characters with the minimum number of character changes (here insertions and deletions). The problem of finding most parsimonious tree, under most parsimony models, is NP-complete [9] and thus the corresponding algorithms are computationally intense. However, a more significant drawback comes from the observation that evolutionary trees constructed with these methods are sometimes incorrect. In this work, we focus on the second problem.

The correctness of the evolutionary tree obtained using a parsimony method depends strongly on the characters used to infer the tree. Intuitively, characters that are easy to gain and easy to lose are not appropriate to use with maximum parsimony methods. Extensive independent acquisition and/or loss of characters in several lineages can make it difficult, if not impossible, to recover the correct evolutionary relationships. At the same time, any realistic approach has to tolerate some events of this type. Therefore, it is important to be able to distinguish characters that provide a consistent evolutionary signal from those which do not. We propose a graph theoretical approach to address this problem.

As mentioned above, we use a particular type of network - a character overlap graph. The vertices of a character overlap graph are characters, and there is an edge between two such characters if and only if there exists a taxon that contains both characters. First, we focus on characters that we call *persistent*. A character is persistent if it is gained exactly once and lost at most once. Thus, the assumption of persistence is weaker than what is required in *perfect parsimony* (where a character can change state only once) but stronger than in Dollo parsimony (where there is no restriction on the number deletions of any given character). We show that a character overlap graph for persistent characters cannot contain network motifs known as *holes*. The simplest hole is a cycle of four nodes with no diagonal edges (*chords*) and is also referred to as a *square*. In general, a *hole* is a chordless cycle of length at least four (Figure 2).

The requirement that all characters be persistent, although weaker than the assumption of perfect parsimony, is still very restrictive. However, the criterion for recognizing persistent characters suggests a heuristic for evaluating whether a given set of characters is hard-to-gain and hard-to-lose in a less restrictive sense. Our simple measure relies on counting squares in the character overlap graph constructed for a given set of characters, and comparing the count to the number of squares expected by chance. (This approach can easily be extended to counting also larger holes, e.g. of size 5, but identifying all holes in a large graph is computationally infeasible.) Furthermore, nodes involved in a large number of squares can be used to identify characters whose removal is likely to improve the results of a parsimony method.

We applied our technique to two types of characters: protein domains and introns. In eukaryotic organisms, most of the proteins are made up of several domains. Domains are conserved evolutionary units that are assumed to fold independently, and are observed in different proteins with different neighboring domains. Introns are non-coding DNA sequences that interrupt the flow of a gene coding sequence in eukaryotic genes. It has been widely accepted that the probability of gaining an intron independently at the same position in two different organisms is relatively low [11]. In terms of introns persisting through the evolution, the picture is mixed. They are remarkably conserved between some lineages (e.g. between Arabidopsis and Human), but they are lost at a significant rate in other organisms (e.g. worm) [27].

We tested a large set of domain overlap graphs and found that squares are significantly under-represented as compared to what is expected by chance. This is in line with the results of Deeds *et al.* [10] and Winstanley *et al.* [29]. They report a successful reconstruction of evolutionary trees using the Dollo parsimony where (structural) domains are taken as characters. In contrast, the intron overlap graph has nearly as many squares as is expected by chance, indicating a very noisy signal. This explains the observation that the tree constructed from intron data using Dollo parsimony method is incorrect [27].

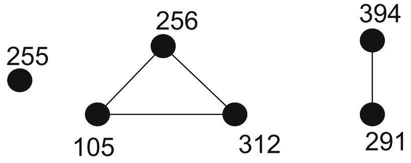
Examining the distribution of squares in each network provides additional insight into the properties of corresponding characters. For both character types, we find that the distribution of squares is non-uniform. For example, in the domain overlap graph, a small number of domains is involved in a large number of holes (see Figure 2). Removal of about 3% of the domains leaves the domain overlap graph square-free. Characteristically, the group of removed domains contains known promiscuous domains (domains known to appear in a large number of diverse proteins). It is indeed appropriate not to include them on equal footing with other characters in parsimony methods.

As mentioned before, the number of squares in the intron overlap graph is very large and it was not clear if removal of the inconsistencies represented by these squares would lead to a meaningful result. We devised a heuristic algorithm to remove squares from the intron overlap graph. Interestingly, we obtained the evolutionary tree consistent with the Coelomata hypothesis [1,2,5,30]. One can think of squares removal as a process that selects a set of characters that are likely to yield a correct tree. This is very much like choosing pairwise compatible characters and building the tree based on these characters alone. However, it is important to point out that, since the concept of persistence is less stringent than pairwise compatibility, this method can be successful when the compatibility method fails. In particular, as shown later in the paper, replacing persistence by pairwise compatibility in the context of intron data does not lead to a correct tree.

2 Characters, Character Overlap Graphs and Parsimony Methods

Characters and parsimony methods. Assume that we are given a set of taxa such that each taxon is characterized by a vector of characters. Intuitively, a character can be anything that describes the properties of a taxa, e.g. external characteristics (like wings, legs, etc.) or a molecular information (like genes, protein domains, etc.). In this work, we assume binary characters, that is, characters that take either value one or value zero (interpreted respectively as the presence/absence of the given characteristics in the taxon). Assume that, during the evolution, characters can be gained and/or lost. Under this assumption, the evolution of a given set of taxa is often reconstructed using parsimony methods. The underlying assumption of parsimony methods is that the characters evolve in a way that minimizes character changes. The maximum parsimony tree is a tree whose leaves are labeled with the character vectors associated with the input taxa, and internal nodes are labeled with the inferred character vectors of ancestral taxa such that the total number of character changes along the tree branches is minimized. Additional restrictions on the type, number, and direction of changes lead to a variety of specific parsimony models [11]. For example, in Dollo parsimony, a character may be inserted (change state from zero to one) only once, but it can be lost multiple times [15]. In Camin-Sokal parsimony, no reversal of character changes is allowed [8]. The problem of computing the maximum parsimony tree is NP-complete for most of parsimony models, including Dollo parsimony and Camin-Sokal parsimony mentioned above [9].

A major problem with parsimony methods is (in addition to their computational cost) that they sometimes produce an obviously incorrect tree. This elucidates the importance of being able to decide if a given character set is likely to be misleading when used in conjunction with a parsimony method. Intuitively, we are interested in characters that are not very easy to gain (thus the number of independent insertions of the same character is limited) and which persist through evolution, i.e. they are not too easy to lose. We propose a graph-theoretical measure that can be used to test whether a given selection of characters is likely to produce the correct evolutionary tree.



	105	255	256	291	312	394
Sc	0	0	0	1	0	0
Dr	1	0	0	0	0	0
Ar	0	0	1	0	0	1
Ce	1	0	1	0	0	0
Hs	1	0	1	0	1	0
Sp	0	1	0	0	0	0
Ag	1	0	0	0	0	0
Pf	0	0	1	0	0	0

Fig. 1. The intron overlap graph for KOG0009 [28]. The introns are identified by the position in the multiple alignment of the corresponding genes. In the matrix on the right side, of the figure rows correspond to the species included in the KOG *Arabidopsis thaliana* (At), *Homo sapiens* (Hs), *C.elegans* (Ce), *Drosophila melanogaster* (Dm), *Anopheles gambaie* (Ag), *Saccharomyces cerevisiae* (Sc), *Schizosaccharomyces pombe* (Sp), and *Plasmodium falciparum* (Pf), and colums correspond to introns identified by Rogozin *et al.* [27] where 1 correspond to the presence and 0 to the absence of the intron at a given position in the multiple alignment.

Character overlap graph. To answer the question whether a given set of characters is hard-to-gain and hard-to-lose, we introduce the concept of a character overlap graph. A character overlap graph is a graph $G = (V, E)$, where V is a set of characters, and $(u, v) \in E$ if there exists a taxon T in the set such that both u and v are present T.

In this paper, we consider two examples of character overlap graphs: a domain overlap graphs and intron overlap graphs. The first family of graphs, also known as domain co-occurrence graphs or domain graphs, has been studied before [31,3,25,32]. A set of taxa used to construct a domain overlap is a family of multidomain proteins. The vertices of the domain overlap graph correspond to protein domains and two domains are connected by an edge if and only if there is a protein that contains both domains. In turn, a set of taxa used in the construction of an intron overlap is a set of completely sequenced genomes. The nodes of an intron overlap graph correspond to the introns and there is an edge between two introns if and only if there is a genome that contains both introns (see Figure 2). No construction equivalent to intron overlap graph has been considered before.

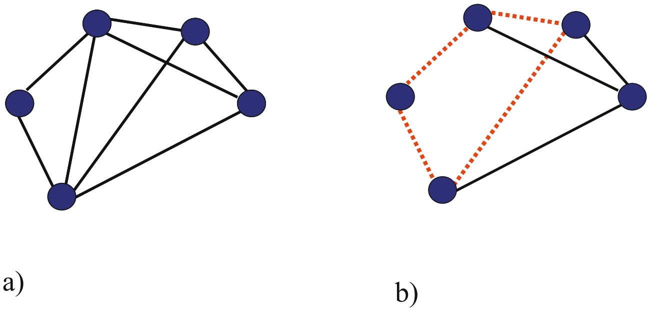


Fig. 2. a) An example of a chordal graph. b) A graph that is not chordal. The red (dotted) circle forms a hole of size four – a square.

Holes and chordal graphs. Chordal graphs constitute a well studied family of graphs [14]. A *chord* in a graph is an edge that connects two non-consecutive vertices of a cycle. A *chordal graph* is a graph which does not have chordless cycles of length greater than three. Chordless cycles of length more than three are called *holes*. Figure 2 (a) shows an example of a chordal graph and Figure 2 (b) a graph which contains a hole of size four – a *square*.

There is a powerful connection between chordal graphs and trees [12,7,19,18, 25], which has been exploited before in the context of phylogenetic trees. We do not use this connection in the paper explicitly, but it is a key result in chordal graph theory and our approach is motivated by this relation.

3 Holes and Parsimony Methods

Graph chordality and persistent characters. We start with an extreme case in which we assume that each character can be gained exactly once and lost at most once. We call such characters *persistent*. Thus a persistent character can undergo at most two changes and these changes are required to respect the order: $0 \to 1 \to 0$. Note that the persistence property is independent of the way a tree is rooted. We show the following simple theorem about persistent characters.

Theorem [characterization of persistent characters]. *If all characters are persistent then the corresponding character overlap graph is chordal.*

Proof. By induction on the size k of the hole.[1] For $k = 4$, assume that there exists a square spanning nodes (characters) A, B, C, and D. This implies that there are four taxa containing respectively pairs of characters AB, BC, CD, and DA, but there does not exist a taxon containing diagonal pairs AC or BD. In fact, no taxon can contain three or more of A, B, C, D simultaneously. Ignoring all other characters, there are, (up to symmetry), two possible binary topologies for the parsimony tree for the four taxa (Figure 3). Since there can be only one insertion per character, all taxa (ancestral or not) containing a specific character must form a connected subtree in the parsimony tree. For example, all nodes on the path from the taxon with characters AB to the taxon with characters BC must contain character B (see Figure 3). Repeating this argument for all pairs of taxa, we infer that the labeling of the two internal nodes in Figure 3 a must contain, respectively, characters A, B, D and B, C, D. By examining all the possibilities it can be seen that this labeling cannot be achieved without deleting at least one character twice. The argument for the case represented in Figure 3 (b) is similar.

Assume now that the graph has a hole $A_0, A_1, \ldots A_{k-1}$ of size k, where $k > 4$. Then for any i there exists a taxon containing the pair of characters A_iA_{i+1} (index additions/subtractions are mod k) but not containing any other A_j where $j \neq i, i+1$. Assume that there exists a parsimony tree T that allows for at most

[1] A shorter proof can be made based on the relation between chordal graphs and trees mentioned in the previous section, but in the interest of keeping the paper self-contained we present here a direct argument.

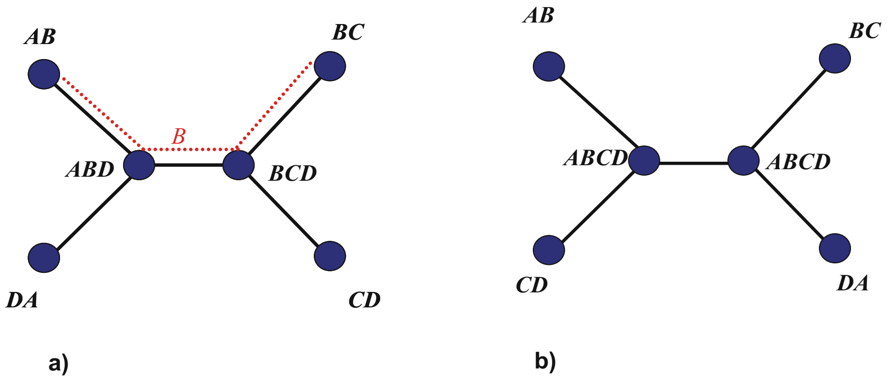
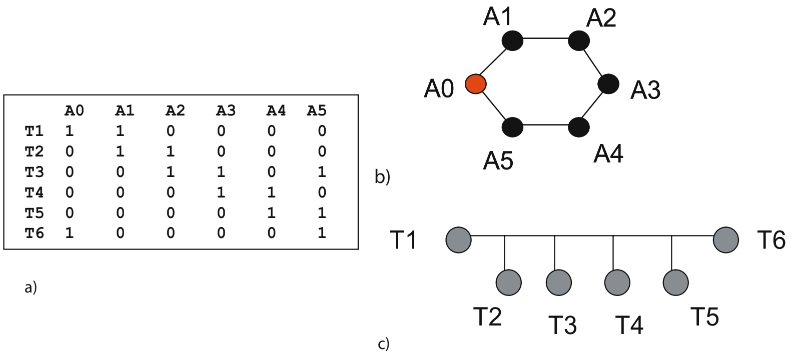


Fig. 3. The two possible (up to symmetry) topologies for an evolutionary tree for four taxa containing respectively characters: AB, BC, CD and DA. Under the assumption that only one insertion per character is allowed, in each case there must exist a character which is deleted twice.

one insertion and one deletion of each character. Consider the subtree of T spanning the taxa involved in the cycle. Let X be an internal node of this tree which is adjacent to two leaves (such node must exist). First, we argue that the two leaves must correspond to two consecutive taxa on the circle. Assume otherwise and let the two leaves be described by character pairs A_iA_{i+1} and A_jA_{j+1} where $j \neq i+1$ and $i \neq j+1$. Then X must contain characters $A_iA_{i+1}A_jA_{j+1}$. (This observation follows from the fact that each of the four characters also belongs to a taxon other than the two taxa corresponding to the leaves A_iA_{i+1} and A_jA_{j+1} and that no double insertions are allowed.) Consider now the subtree spanned by the leaves A_iA_{i+1}, A_jA_{j+1}, the internal node X, and the leaves containing characters $A_i, A_{i+1}, A_j, A_{j+1}$ other than the leaves corresponding to the pairs A_iA_{i+1} and A_jA_{j+1}. By a case analysis similar to the one for the base case if find that the topology of this tree contradicts the assumption of single insertion/deletion. Thus the two leaves must correspond to two consecutive taxa in the circle, that is without loss of generality, $A_{i+1} = A_j$. Now we are ready to use the inductive hypothesis. Replace the pair of taxa with characters A_iA_{i+1} and $A_{i+1}A_{i+2}$ with one taxon with characters $A_iA_{i+1}A_{i+2}$ and consider the tree T' obtained from the tree T by removing leaves corresponding to A_iA_{i+1} and $A_{i+1}A_{i+2}$. If T is a tree that does not require more than one insertion and more than one deletion per character so is T' with respect to the modified set of taxa. By the inductive hypothesis, this is impossible since the character overlap graph for the reduced set of taxa contains a cycle of size $k-1$. QED.

Persistence versus Compatibility. The persistence criterion provided above is similar to the well known compatibility criterion [11] at the basic level. Namely, they both seek to identify characters that are in some sense inconsistent. Then, one can look for a set of characters whose removal leaves a set of consistent characters and construct the tree based on these consistent characters. There are, however, important differences. In the case of persistent characters, a character



	A0	A1	A2	A3	A4	A5
T1	1	1	0	0	0	0
T2	0	1	1	0	0	0
T3	0	0	1	1	0	1
T4	0	0	0	1	1	0
T5	0	0	0	0	1	1
T6	1	0	0	0	0	1

Fig. 4. a) An artificial example of 6 taxa and 6 characters; (b) corresponding character overlap graph; c)the most parsimonious tree after removing character A_0. Note that in this case half of the characters would have to be removed to obtain a pairwise compatible set.

can change the state at most twice (one insertion and at most one deletion) while for pairwise compatible characters each character changes state at most once. Thus the assumption of persistence is a weaker assumption than that of compatibility. In particular it is easy to see that every edge in a hole identifies a pair of non-compatible characters. Consider for example a set of n taxa each described by two characters $A_0A_1, A_1A_2, \ldots A_nA_0$. Then characters A_iA_{i+1} are incompatible. Removing just one character will ensure persistence (and later in this paper we propose a method to decide which one) while one has to remove half of the characters to obtain pairwise compatible set (see figure 4). This weaker consistency requirement is particularly useful when one cannot assume that there exist a sufficiently large set of characters which once inserted are never lost. An example of such situation occurs in the case of intron evolution.

Finally, we shall point out that, unlike the compatibility criterion, the theorem shown provides a necessary but not sufficient condition for persistence of characters.

Graph motifs and persistent characters. The requirement that each character must be persistent is very restrictive. For example, the fact that bats and birds gained wings independently (that is the character wings is gained twice) does not lead to an incorrect evolutionary tree as long as other characters are used to complement this information. So, even if characters are occasionally gained/lost more than once, we may still be able to apply parsimony methods successfully. However, if characters are gained and/or lost independently on massive scale, then there is not much hope of recovering the correct tree. How can one distinguish between these two cases?

One solution would be to measure how far the corresponding character overlap graph is from being a chordal graph. This can be measured, for example, by counting the minimal number of edges whose addition makes the graph chordal. Unfortunately, the problem of finding such a minimal set is NP-complete [17].

We propose a simple heuristic based on the network motifs approach. Rather than considering all holes, we consider only holes of size four (squares). All squares can be easily enumerated. The number of squares in an attribute overlap graph can then be compared to the number of squares in a null model, where the characters are gained/lost randomly. The ratio of these two counts can be used to measure how easily the characters are gained/lost.

Our null model assumes the same number of taxa as the real data and the same set of characters. Furthermore, there is a one-to-one correspondence between the real taxa and the taxa in the null model. In the null model, the characters of each taxon are selected randomly in a way that for each taxon the expected number of characters equals the number of characters of the corresponding real taxon.

In general, each square in a character overlap graph indicates existence of a non-persistent character. While a small number of squares is clearly an indication of a persistent nature of most of the characters, the large number of squares does not necessarily indicate that the number of non-persistent characters is equally large. Squares can overlap and a small number of non-persistent characters can result in a relatively large number of squares. Thus, characters involved in a large number of squares introduce significant noise to the data. One can address this problem by assigning a smaller weight to these characters, or simply by removing them from the data.

4 Applications and Experimental Results

Construction of intron overlap graph and domain overlap graphs. To construct intron overlap graph, we used the data from a study by Rogozin *et al.* [27]. This data contains information about introns found in conserved (and orthologous) genes of eight fully sequenced organisms: *Arabidopsis thaliana* (At), *Homo sapiens* (Hs), *C.elegans* (Ce), *Drosophila melanogaster* (Dm), *Anopheles gambaie* (Ag), *Saccharomyces cerevisiae* (Sc), *Schizosaccharomyces pombe* (Sp), and *Plasmodium falciparum* (Pf). Introns are identified by their starting position with respect to the coding sequence. The data contains information about 7236 introns, however most of these introns are observed in one organism only. After eliminating these single-organism entries, we were left with 1790 introns.

To construct domain overlap graphs, we used the data from a study by Przytycka *et al.* [25] containing 479 multi-domain superfamilies. This data was built using the non-redundant multidomain proteins in Swiss-prot [6], where the domains were recognized using CDART [13]. Proteins in this set are grouped into overlapping superfamilies. Each superfamily is defined to be the maximal set of proteins that have a specific domain in common. For example, all proteins containing the kinase domain form one superfamily, proteins containing the SH2 domain form another superfamily and these two superfamilies intersect. Each such superfamily is considered to have its own evolutionary history, therefore, each superfamily is treated separately. For each such superfamily there is a separate domain overlap graph. Domain overlap graphs with less than four nodes

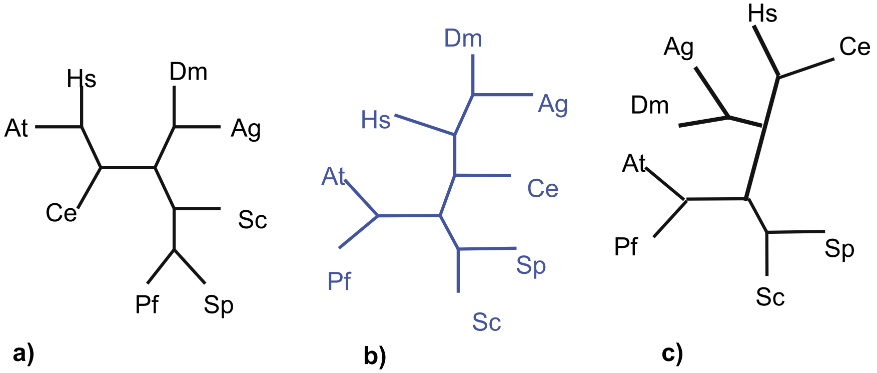


Fig. 5. Three tree topologies for organisms: *Arabidopsis thaliana* (At), *Homo sapiens* (Hs), *C.elegans* (Ce), *Drosophila melanogaster* (Dm), *Anopheles gambaie* (Ag), *Saccharomyces cerevisiae* (Sc), *Schizosaccharomyces pombe* (Sp), and *Plasmodium falciparum* (Pf) a) The incorrect Dollo parsimony tree computed from intron data b) The tree consistent with Coelomata hypothesis. This is also exactly the tree obtained after applying our squares removal procedure. c) The tree consistent with Ecdysozoa hypothesis.

Table 1. The frequencies of occurrences of squares in intron overlap graph and domain overlap graph relative to the corresponding null random model. Observe significant under-representation of squares in the domain overlap graph.

Character type	# squares in character overlap graph (s)	# squares expected by chance
Introns	954 667 368	1 389 751 510
Domains	251	3 822

were ignored. Similarly, networks in which the number of edges was smaller than the number of nodes were disregarded.

Counting squares. The relative numbers of squares for both types of overlap graphs are summarized in Table 1. In the domain overlap graphs, the total number of squares is relatively small. This indicates that domains tend to be persistent and thus provide a good set of characters to be used by parsimony methods. In contrast, the intron overlap graph contains nearly as many squares as it is expected by chance. This suggests that applying parsimony methods to this data is likely to give an incorrect result. Indeed, Rogozin *et al.* constructed such tree (using Dollo parsimony) and found that it is completely wrong. Figure 5 shows the result of this construction, Figure 5 (b) the tree consistent with the Coelomata hypothesis, and Figure 5 (c) the tree consistent with the Ecdysozoa hypothesis. Interestingly, the incorrect Dollo tree is supported by high bootstrap values [27] suggesting that the incorrectness of the tree is due to a systematic bias rather than a random noise.

Eliminating squares in domain overlap graphs. Figure 6 shows the distribution of number squares summed up over all domain overlap graphs. We observe that a few domains are involved in a large number of squares. Since the

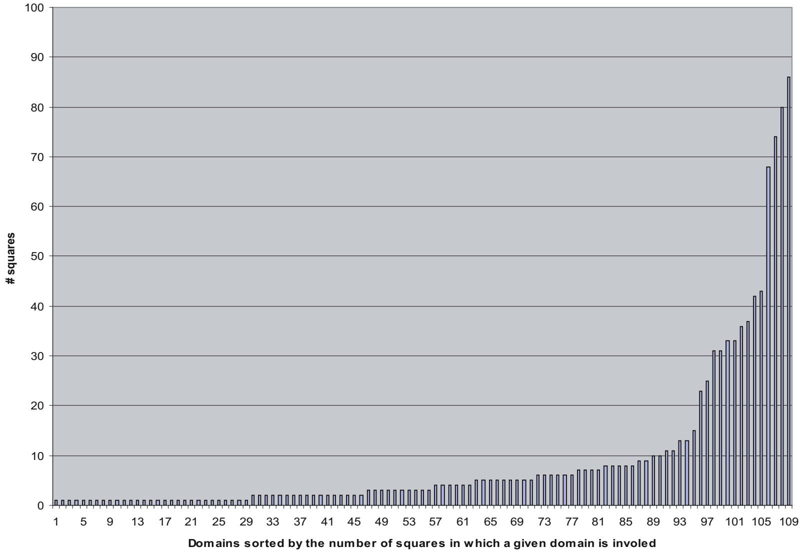


Fig. 6. The distribution of squares. The domains (x axis) are sorted in the increasing order of the number of squares they belong to.

problem of removing the smallest number of nodes to obtain a hole-free graph is NP-complete ([33]), we iteratively removed the node involved in the highest number of squares (re-computing the number of squares each time). The first two domains removed by our greedy approach are two functionally uncharacterized domains (smart00267 and smart00052 [16]). Subsequently, the algorithm identifies for removal two known promiscuous domains (domains that are known to appear in many diverse multidomain proteins): SH2, ABC-ATPase. Removal of these four domains reduces the already small number of squares by nearly 80%. After this step, there are still a few domains involved in squares including: PDZ, PH, SH3, EGF, IG-like. However, none of these domains are involved in more than 11 squares.

Eliminating squares in the intron overlap graph. In the case of intron overlap graph, the number of squares is not much smaller than what is expected by chance. The most frequently occurring squares are of type: (At, Hs, Ce, X), where X is Dm or Ag. Note that each intron is represented by a binary pattern of length eight (the number of genomes in the data) where one corresponds to the intron being present in the given genome and zero to its absence. Introns with the same pattern are indistinguishable from the perspective of parsimony methods and are involved in the same number of squares. Note further that with eight species there are $2^8 - 9$ intron patterns (the subtraction corresponds to the assumption that each intron must be in at least two species) out of which 90 patterns were populated. Thus, some patterns are represented multiple times. The patterns that appear significantly more often than it is expected by chance are considered to be more informative (more significant). Let n_i be the number of times pattern i is observed in the intron data, and r_i expected number of occurrences of the pattern in the null model. Define $p_i = \frac{n_i}{r_i}$ to be the significance the intron pattern i. (Using $p_i = \max(\log \frac{n_i}{r_i}, \epsilon)$, where ϵ is a real number closed to zero (here $= 10^{-10}$) gave the same results.) Let S_i be the number of squares in which an intron with pattern i is involved. Our greedy square removal algorithm removes iteratively intron patterns that maximize the value $\frac{S_i}{p_i}$. This provides a trade off between maximizing the number of removed squares and minimizing the significance of the removed intron patterns. After all squares are removed, we apply the Dollo parsimony to the remaining introns. The procedure removed intron 52 (57 %) patterns. We also introduce a modification to the Dollo parsimony with enforces that the contribution of each intron is weighted with the significance of the corresponding intron pattern. The resulting tree is presented on Figure 5 (b). Thus we obtained a tree which is consistent with the Coelomata hypothesis.

We also applied the same greedy approach with the persistence criterion replaced with the compatibility criterion. The procedure removed 86 (95 %) of intron patterns and produced 15 incorrect trees.

5 Discussion, Conclusions, and Further Work

We demonstrated that the character overlap graph for persistent characters is chordal. This suggests that the character overlap graph for characters that are hard-to-gain and hard-to-lose is expected to contain relatively few holes as compared to a null model. In particular, the number of holes of size four (squares) is also expected to be relatively small. The last property is easily testable, and provides a fast method for checking whether a set of characters can be used to produce a correct evolutionary tree. In practical applications, we found that the number of squares in the domain overlap graph is very small, supporting the findings that domains can be used as characters in a parsimony approach. In contrast, the number of squares in the intron overlap graph is not much smaller than it is expected by chance. This explains why the Dollo tree built based on intron data is incorrect.

A large number of squares does not necessarily indicate that all characters are non-persistent. For example, we demonstrated that in the domain overlap graph, the majority of squares come from the existence of a handful of promiscuous domains. Consequently, removing a small number of domains from this character set leaves the domain overlap graph square free.

A similar approach applied to the intron overlap graph also produced an interesting result. While it is known that introns can be remarkably conserved in some lineages, they are not so conserved in others. This leads to a large number of squares. However, we found that the distribution of these squares is non-uniform. Thinking of squares as inconsistencies in the data, we applied a greedy algorithm to remove introns that are involved in square formation, choosing introns of low significance and high involvement in square motifs. We used this truncated intron data to construct a weighted Dollo parsimony tree. That is, we weighted the contribution of each intron according to the significance of the corresponding intron pattern. With these two changes, we obtained a parsimony tree which is consistent with the tree constructed using other methods [30]. This is in contrast to the previous applications of parsimony methods, which have been unable to recover a tree consistent with any of the proposed evolutionary hypotheses.

The results of this work strongly suggest that removal of non-persistent characters involved in a large number of squares may significantly improve the applicability of parsimony methods.

Acknowledgments. All algorithms have been implemented with the help of LEDA tools [20]. The author thanks David Lipman (NCBI), Igor Rogozin (NCBI), Uri Wolf (NCBI) for stimulating discussions, and Raja Jothi (NCBI) and Elena Zotenko (NCBI/UMD) for discussions and critical comments on the manuscript. This research was supported by the intramural research program of the National Institutes of Health.

References

[1] A. Adoutte, G. Balavoine, N. Lartillot, O. Lespinet, Benjamin Prud'homme, and Renaud de Rosa. Special Feature: The new animal phylogeny: Reliability and implications. *PNAS*, 97(9):4453–4456, 2000.

[2] A.M. Aguinaldo, J.M. Turbeville, L.S. Linford, M.C. Rivera, J.R. Garey, R.A. Raff, and J.A. Lake. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387:489–93, 1997.

[3] 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.

[4] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[5] Jaime Blair, Kazuho Ikeo, Takashi Gojobori, and S Blair Hedges. The evolutionary position of nematodes. *BMC Evolutionary Biology*, 2(1):7, 2002.

[6] B. 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] P. Buneman. A characterisation of rigid circuit graphs. *Discrete Math.*, 9:205–212, 1974.

[8] J. H. Camin and R.R. Sokal. A method for deducting branching sequences in phylogeny. *Evolution*, 19:311–326, 1965.

[9] W.H.E. Day, D. Johnson, and D. Sankoff. The computational complexity of inferring rooted phylogenies by parsimony. *Mathematical Biosciences*, 81:33–42, 1986.

[10] E.J. Deeds, Hooman Hennessey, and Eugene I. Shakhnovich. Prokaryotic phylogenies inferred from protein structural domains. *Genome Res.*, 15(3):393–402, 2005.

[11] J. Felsenstein. *Inferring Phylogenies.* Sinauer Associates, 2004.

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

[13] 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.

[14] M. Golumbic. *Algorithmic Graph Theory and Perfect Graphs.* Academic Press, New York, 1980.

[15] J.S.Farris. Phylogenetic analysis under Dollo's law. *Systematic Zoology*, 26(1):77–88, 1977.

[16] I. 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.

[17] J. M. Lewis and M. Yannakakis. The node-deletion problem for hereditary properties is NP- complete. *J. Comput. Syst. Sci.*, 20(2):219–230, 1980.

[18] T.A. McKee and F.R. McMorris. *Topics in intersection graph theory.* SIAM Monographs on Discrete Mathematics and Applications, 1999.

[19] F.R. McMorris, T. Warnow, and T. Wimer. Triangulating vertex colored graphs. *SIAM J. on Discrete Mathematics*, 7(2):296–306, 1994.

[20] K. Mehlhorn and S. Naher. *The LEDA Platform of Combinatorial and Geometric Computing.* Cambridge University Press, 1999.

[21] M. Middendorf, E. Ziv, and C. H. Wiggins. From The Cover: Inferring network mechanisms: The Drosophila melanogaster protein interaction network. *PNAS*, 102(9):3192–3197, 2005.

[22] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of Evolved and Designed Networks. *Science*, 303(5663):1538–1542, 2004.

[23] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827, 2002.

[24] N. Przulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[25] T.M. Przytycka, G. Davis, N. Song, and D. Durand. Graph theoretical insight into evolution of multidomain proteins. *Lecture Notes in Computational Biology (RECOMB 2005)*, 3500:311321, 2005.

[26] T.M. Przytycka and Y.K. Yu. Scale-free networks versus evolutionary drift. *Computational Biology and Chemistry*, 28:257–264, 2004.

[27] I.B. Rogozin, I.Y Wolf, A.V. Sorokin, B.G. Mirkin, , and V Koonin, E. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Current Biology*, 13:1512–1517, 2003.

[28] R. Tatusov, N. Fedorova, J. Jackson, A. Jacobs, B. Kiryutin, E. Koonin, D. Krylov, R. Mazumder, S. Mekhedov, A. Nikolskaya, B.S. Rao, S. Smirnov, A. Sverdlov, S. Vasudevan, Y. Wolf, J. Yin, and D. Natale. The cog database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1):41, 2003.

[29] Henry F. Winstanley, Sanne Abeln, and Charlotte M. Deane. How old is your fold? *Bioinformatics*, 21(suppl1):i449–458, 2005.

[30] Y.I. Wolf, I.B. Rogozin, and E.V. Koonin. Coelomata and Not Ecdysozoa: Evidence From Genome-Wide Phylogenetic Analysis. *Genome Res.*, 14(1):29–36, 2004.

[31] S. Wuchty. Scale-free behavior in protein domain networks. *Mol. Biol. Evol.*, 18:1694–1702, 2001.

[32] S. Wuchty and E. Almaas. Evolutionary cores of domain co-occurrence networks. *BMC Evolutionary Biology*, 5(1):24, 2005.

[33] M. Yannakakis. Computing the minimum fill-in is NP- complete. *SIAM J. Alg and Discrete Math*, 2:77–79, 1981.

[34] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, and H. Margalit. Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *PNAS*, 101(16):5934–5939, 2004.