

Obtaining Maximal Concatenated Phylogenetic Data Sets from Large Sequence Databases

Michael J. Sanderson,* Amy C. Driskell,* Richard H. Ree,† Oliver Eulenstein,‡ and Sasha Langley*

*Section of Evolution and Ecology, University of California, Davis; †Botanical Garden and Centre for Plant Research, University of British Columbia; and ‡Department of Computer Science, Iowa State University

To improve the accuracy of tree reconstruction, phylogeneticists are extracting increasingly large multigene data sets from sequence databases. Determining whether a database contains at least k genes sampled from at least m species is an NP -complete problem. However, the skewed distribution of sequences in these databases permits all such data sets to be obtained in reasonable computing times even for large numbers of sequences. We developed an exact algorithm for obtaining the largest multigene data sets from a collection of sequences. The algorithm was then tested on a set of 100,000 protein sequences of green plants and used to identify the largest multigene ortholog data sets having at least 3 genes and 6 species. The distribution of sizes of these data sets forms a hollow curve, and the largest are surprisingly small, ranging from 62 genes by 6 species, to 3 genes by 65 species, with more symmetrical data sets of around 15 taxa by 15 genes. These upper bounds to sequence concatenation have important implications for building the tree of life from large sequence databases.

Introduction

Improved accuracy of phylogenetic inference through the concatenation of multiple sequences from the same taxon is expected on theoretical grounds (Erdos et al. 1999; Bininda-Emonds et al. 2001) and has been found in many recent studies (Qiu et al. 1999; Soltis, Soltis, and Chase 1999; Graham and Olmstead 2000; Brown et al. 2001; Madsen et al. 2001; Murphy et al. 2001; Baptiste et al. 2002). For example, Baptiste et al. (2002) combined 123 genes for 30 species of eukaryotes and found support increased monotonically with the number of genes included. Concatenation has been undertaken widely in taxa in which complete genomes are available, such as in prokaryotes (Brown et al. 2001); taxa having small organellar genomes, such as animal mitochondria (Miyata and Nishida 2000); and taxa for which long-term coordination among investigators has driven parallel sequencing efforts, as for green plants (Chase et al. 1993). Adding genes for a given set of taxa generally improves both robustness and running times of computationally intensive phylogenetic analyses (Soltis et al. 1998; Savolainen et al. 2000; Baptiste et al. 2002), although inferences about any particular node may sometimes be skewed by long-branch attraction (Felsenstein 1978). Adding taxa for a given set of genes can also sometimes improve phylogenetic tree inference, especially in cases of long-branch attraction (Hillis 1996). Increased taxon sampling can improve reconstructions of ancestral character states, sharpen estimates of rates and modes of sequence evolution, and generally provide a more comprehensive summary of a clade's history. Not surprisingly, the dimensions of phylogenetic sequence data sets have grown rapidly in recent years.

However, phylogenetic methods require input data in the form of rectangular matrices of taxa by aligned sequences. Ideally, such data sets are *complete*—meaning that every species has been sequenced for every gene in

the data matrix. The sample of genes among taxa found in sequence databases or available from other sources rarely allows large complete data sets to be constructed, however, and therefore all large concatenated phylogenetic data sets published recently have missing entries (Qiu et al. 1999; Soltis, Soltis, and Chase 1999; Murphy et al. 2001). Missing data in incomplete data sets should eventually degrade phylogenetic inference by increasing both the number of optimal solutions found and the uncertainty in the placement of some taxa relative to others (Wiens 1998), although how much missing data is tolerable remains an open question (Kearney 2002). This motivates the present study. How can we optimally construct complete phylogenetic data sets from large sequence databases? The problem can be posed more formally as follows: Given a large collection of sequences that have been partitioned into sets of homologous genes, is it possible to construct a complete data matrix in which m taxa have sequences for the same k genes? Moreover, is it possible to find all complete data sets of this size or larger? Finding complete data sets in a sequence database is a nontrivial problem. In fact, determining whether a complete data set of a given size exists is an NP -complete problem, meaning that efficient (polynomial time) algorithmic solutions are unlikely to be discovered (Garey and Johnson 1979). However, we describe an exact, exponential time algorithm which effectively solves many large problems, and we illustrate its use by constructing maximal concatenated data sets from a large set of orthologous protein sequences available from green plants. Alexe et al. (2002) have described a different algorithm which, though polynomial in the output size, also has worst-case exponential running time, because the output size can grow exponentially with input size.

Materials and Methods

Definitions

A *cluster* is a set of sequence homologs. Because our sequences consist entirely of protein coding genes, a cluster here represents a “gene” or “protein.” Concatenation of sequences is appropriate only for clusters consisting of orthologous genes, so we restrict attention to those (see

Key words: biclique, NP -complete, sequence concatenation, phylogeny, optimization.

E-mail: mjsanderson@ucdavis.edu.

Mol. Biol. Evol. 20(7):1036–1042. 2003

DOI: 10.1093/molbev/msg115

© 2003 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Table 1
Distribution of Sequences and Taxa Among
Phylogenetically Informative Green Plant Clusters

	All Clusters	Informative Clusters	Orthologous Informative Clusters
Number of clusters	40,154	1092 (3%)	656 (2%)
Number of sequences	100,813	38,061 (38%)	22,268 (22%)
Number of taxa	11,527	11,052 (96%)	10,141 (88%)

below and table 1). The *cluster set*, C , is the set of all clusters. A cluster set, C , can be represented as a bipartite graph (West 2001), $G(C)$, consisting of two disjoint sets of nodes, one for the clusters and one for the taxa (fig. 1). Edges connect a cluster node to a taxon node if and only if that cluster has a sequence for that taxon. A *complete* phylogenetic data set is one containing no missing genes for a given set of taxa. Finding a complete phylogenetic data set for a cluster set C is equivalent to finding a *biclique* in the bipartite graph, $G(C)$. A biclique is a subgraph of $G(C)$ denoted $K_{k,m}$ and consisting of k clusters and m taxa such that every cluster node is connected (adjacent) to every taxon node and vice versa (fig. 1). A biclique is *maximal* if no other biclique exists that contains it as a proper subgraph. There may be many maximal bicliques. Biclques arise in many problems in graph theory, and the complexity of problems related to bicliques has been studied in some detail (Hochbaum 1998; Peeters 2000; Dawande et al. 2001).

Formal Statement of Problem

We study the following optimization problem:

Given: Bipartite graph $G(C)$ for cluster set C , natural numbers k and m .

Find: All maximal bicliques, $K_{k',m'}$, for $G(C)$, in which $k' \geq k$, $m' \geq m$.

The decision problem asking whether a biclique, $K_{k,m}$, exists in $G(C)$ is *NP*-complete. This follows because the decision problem version of our problem for the instance $k = m$ is called GT24 (balanced complete bipartite subgraph problem) in Garey and Johnson (1979; see also Peeters 2000) and is known to be *NP*-complete. Therefore the decision variant of our problem (without the restriction $k = m$) must therefore be *NP*-complete. This implies that the optimization problem stated above is also unlikely to have an efficient solution for arbitrary inputs (Cormen et al. 2001).

Algorithm

We have developed an exact algorithm that solves this problem quickly (in minutes to a few hours on a Linux workstation) for our data, by exhaustively building progressively larger bicliques. Briefly, suppose we seek all bicliques at least as large as $K_{k,m}$. The exact algorithm examines every pair of clusters and calculates the intersection of their taxon sets. Let the size of that intersection (the number of taxa in both clusters) be m' . This step finds all bicliques, $K_{2,m'}$. Discard any bicliques in

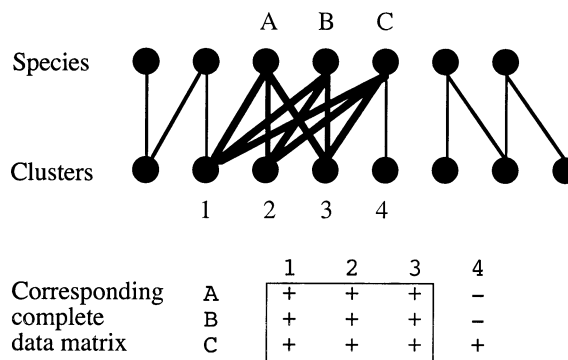


FIG. 1.—Bipartite graph of a hypothetical cluster set. Edges connect species (taxon) nodes with cluster nodes if a sequence exists for that taxon in that cluster. A hypothetical biclique is indicated by bold lines. It corresponds to a phylogenetic data matrix that is complete (shown in box). Cluster 4 has not been sequenced for species A or B and therefore cannot be part of this complete matrix.

which $m' < m$, and—assuming any are left—try adding every remaining cluster to each surviving candidate biclique. This step finds all bicliques, $K_{3,m'}$. If this iterative procedure terminates before finding bicliques, $K_{k,m}$, then none exists satisfying the constraints. On the other hand if the iteration gets to the point of examining k clusters, then every biclique found from then on will be of the required size or larger, and the algorithm will continue to enumerate these until it has found all of them. Maximal bicliques are obtained from these by discarding any that are contained in another biclique. This algorithm was implemented in C++, using the LEDA library (Algorithmic Solutions Software GmbH) for data structures. Source code and a Linux executable are available at (<http://ginger.ucdavis.edu/sandlab/SOFTWARE>). See the *Appendix* for a detailed description of the algorithm.

Sequence Data, Extraction of Cluster Set, and Identification of Ortholog Clusters

To illustrate the effectiveness of this algorithm, we applied it to a set of 100,813 proteins sampled from 11,587 taxa of green plants extracted from the five GenBank GBPLN flatfiles in release 127.0 of GenBank (December 2001), representing all green plant protein coding sequences in the database that have been translated. The cluster set was constructed using all-against-all BLAST searches (Altschul et al. 1990) combined with single linkage clustering as implemented in the National Center for Biotechnology Information (NCBI) *blastclust* tool (Dondoshansky 2002), at a stringency of 60% at the amino acid level. Other clustering strategies have been described which are considerably more sophisticated than ours (Krause, Stoye, and Vingron 2000; Kriventseva et al. 2001; Tatusov et al. 2001), but our algorithm can be applied to any cluster set, regardless of how it is constructed. All clusters are available in our online MySQL database (host: ginger.ucdavis.edu; user name “guest”).

Sequence concatenation is most appropriate for orthologous sequences. Automated identification of orthologs in sequence database is a challenging task both for

computational reasons (Remm et al. 2001; Lee et al. 2002) and because sparse sampling of gene families in databases imposes sharp limits on the strength of inferences. We used a phylogenetic procedure, rather than relying on more standard reciprocal Blast searching strategies (e.g., Lee et al. 2002). By default, the sequences in any cluster in which only a single sequence was present per taxon were treated as orthologous. For clusters containing multiple sequences in at least one taxon, unconstrained and constrained gene trees were constructed using maximum parsimony with a “protein parsimony” step matrix (Swofford et al. 1996; gaps treated as missing data) in the program PAUP* (Swofford 2002). The constrained tree forced all sequences from the same species to form a clade. If the constrained tree was significantly less parsimonious than the unconstrained tree (using a signed-rank test; Swofford et al. 1996), the cluster was considered to contain paralogs and was excluded from concatenation analysis. Note that a cluster might well contain only the orthologs of one paralog in a gene family. Also, some orthologous genes with ancestral polymorphisms will be excluded by the test, and if sampling of a gene family is extremely poor, such that only one sequence per taxon is found in the database, then we are left with no choice but to mistakenly infer orthology. This will be especially likely for clusters with only a few sequences and taxa.

Phylogenetic Analysis of Concatenated Data Sets

Phylogenies were constructed using standard methods for two representative maximal bicliques selected from the set of maximal bicliques. Clusters containing multiple accessions from the same species were pruned such that only one sequence was included per species. This pruning is justified because all sequences from the same species form a clade for those clusters passing the phylogenetic test for orthology. Presumably these represent multiple accessions of the same gene or multiple alleles from the same locus. Amino acid sequences were aligned with default options in ClustalW (Thompson et al. 1994). Protein parsimony was used to reconstruct trees (see above). Bootstrap analysis (100 replicates) was used to assess phylogenetic support for clades. The single-celled streptophyte *Mesostigma* was used as the outgroup to land plants for the $K_{39,10}$ biclique; three chlorophytes were used as outgroups to the streptophytes (including land plants) in the $K_{15,15}$ biclique. Aligned data sets are available at http://ginger.ucdavis.edu/sandlab/WWW_DATA.

Results

Clustering the database generated a set of 40,154 protein clusters, of which 1,092 were potentially phylogenetically informative about species relationships because they contained four or more distinct taxa (table 1). The plastid genes *rbcL*, *matK*, and *ndhF* formed the largest clusters, and the model organisms *Arabidopsis*, rice, and maize were represented in the most clusters. Of the informative clusters, 656 were determined to consist entirely of orthologs, and thus represent candidates for concatenation. Although this usable subset contained only 26% of the

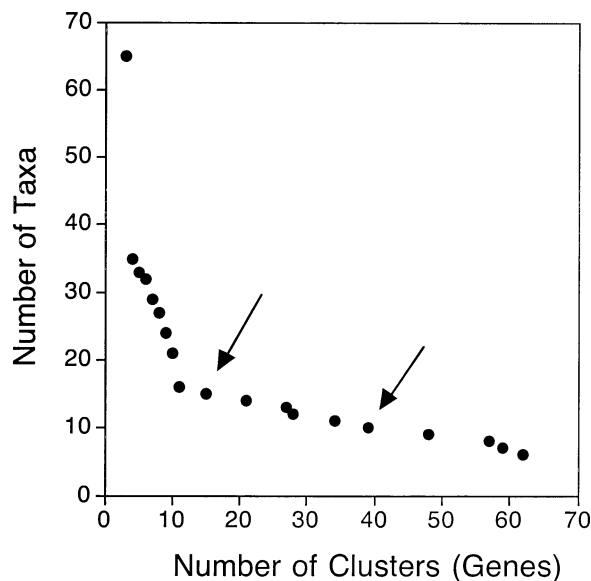


FIG. 2.—Largest maximal bicliques for the green plant protein data set having at least 3 genes (clusters) and 6 species. No bicliques exist above and to the right of those indicated. Every possible data set below and to the left does exist, but many are trivial subsets of the indicated maximal bicliques. The two bicliques subjected to phylogenetic analysis are indicated by arrows.

sequences in the original data set, its taxonomic coverage was still 88% (table 1).

Although the running time of the exact algorithm precluded identification of *all* maximal bicliques, careful setting of different combinations of lower bounds, k , and m , allowed us to quickly identify the *largest* ones (fig. 2). Running times increase rapidly as the input constraints, k and m , are decreased, which is expected given the computational complexity of the problem. To find the largest maximal bicliques (those forming the boundary in fig. 2), it suffices to choose values of k and m that are just small enough to find some bicliques but not so small that run times become a problem. Each such successful run identifies one or more bicliques guaranteed to have the property that no other bicliques exist above and to the right of it (i.e., with larger values of either k or m or both). Ten runs of the algorithm were sufficient to identify all largest maximal bicliques with more than 3 clusters and 6 taxa.

The relative efficiency of the exact algorithm even for this large sequence set stems from the uneven distribution of sequences of genes among taxa (many sequences are available for a few model species and a few heavily sampled genes are available for many species). No bicliques exist larger than the set of maximal bicliques shown, which form a boundary in the space of concatenated data sets (fig. 2). Bicliques of every possible dimension exist below and to the left of the boundary, but many of these are not maximal. Some bicliques are equal to each other in their dimensions but have different taxon and cluster compositions. Many of the bicliques overlap, and for any given biclique there are generally bicliques which have either slightly more taxa and slightly fewer clusters, or slightly more clusters and fewer taxa. The largest bicliques

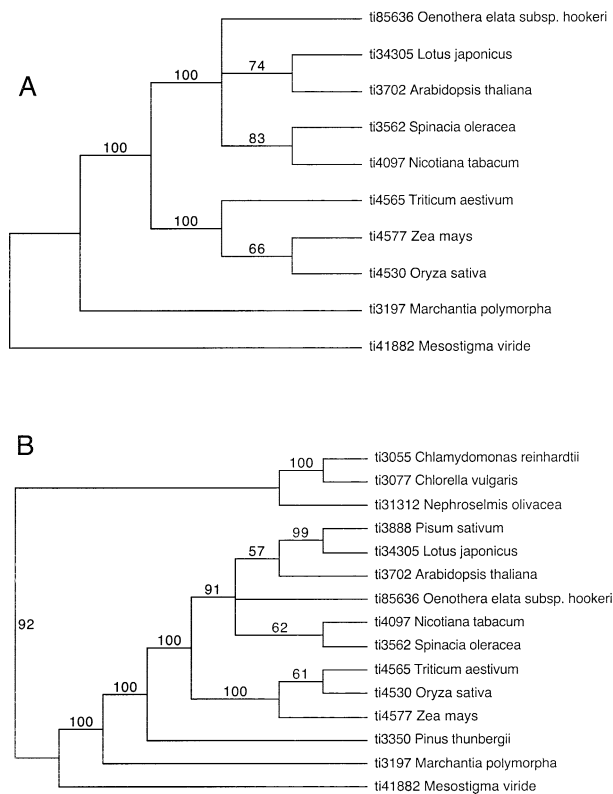


FIG. 3.—A, phylogenetic tree for the $K_{39,10}$ concatenated data set. B, phylogenetic tree for the $K_{15,15}$ concatenated data set. Bootstrap support is indicated at each node. Taxon names are prefixed with their NCBI taxonomy identification numbers.

form a highly left-skewed “hollow” curve with many data sets of few genes and many taxa, or few taxa and many genes. At one end of this range was a data set with 62 genes by 6 species; at the other end, a data set with 3 genes by 65 species. Biclques with approximately equal numbers of both taxa and genes have the unexpectedly small size of about 12–15 of each. In general, no concatenated data set contained more than about 400 sequences. Thus, no complete data set contained more than about 2% of the 22,000 sequences found in phylogenetically informative ortholog clusters for green plants.

The phylogenetic tree (fig. 3) based on the $K_{39,10}$ biclique (39 genes; 8,523 amino acids; 10 species) provides strong support for major clades within land plants, including basal relationships within eudicot angiosperms that are not well resolved in recent analyses (Qiu et al. 1999; Soltis et al. 1999; Savolainen et al. 2000). For example, spinach (*Chenopodiaceae*) and tobacco (*Solanaceae*) share a more recent common ancestor than either does with the other angiosperms sampled. The $K_{15,15}$ biclique (15 genes; 3,919 amino acids; 15 species) samples many fewer genes but delves more deeply into green plant phylogeny, including several green algal relatives of land plants, another angiosperm, and another seed plant, *Pinus*. Bootstrap support was slightly lower within eudicots because of fewer characters, which is apparently the trade-off for increased taxonomic coverage. These examples are

characteristic of the high degree of overlap between many maximal biclques.

Discussion

The algorithms described here permit construction of maximal complete concatenated sequence data sets from the sequence databases. Construction of complete data matrices via concatenation for phylogenetic analysis, equivalent to the identification of biclques, is a computationally hard problem in theory. However, for problem instances such as the one examined above, comprising a set of over 20,000 proteins for a taxonomically diverse collection of green plants (some 10% of all species in GenBank), exact algorithms can solve the problem fairly quickly. This should permit more intensive exploitation of sequence databases for phylogenetic purposes.

Assembling the Tree of Life: The Role of Sequence Concatenation

These results have implications for recent efforts aimed at assembling large parts of the tree of life (<http://research.amnh.org/biodiversity/features/feat.html>; or more correctly, that part of the history of life that is tree-like, see Wolf et al. 2002). Exploiting the size and diversity of sequence databases for building comprehensive species phylogenies poses many computational challenges. Two competing strategies have been discussed: (1) concatenation of sequences for increasingly large sets of taxa, as discussed here; and (2) combination of trees (rather than data) constructed from separate but overlapping data sets, using “supertree” methods (Sanderson, Purvis, and Henze 1998; Daubin, Gouy, and Perriere 2001; Liu et al. 2001). Given that some 11,000 species of green plants have protein sequences in GenBank, and that these fall into 40,000 protein clusters, the discovery that the corresponding maximal biclques have sizes on the order of only 15 by 15 is striking. Moreover, the fact that these complete data sets each contain no more than about 2% of the available sequences implies some limitations on the utility of concatenated data sets. Even though the algorithm described here will permit more intensive exploitation of the sequence databases, concatenation alone will probably not provide a comprehensive solution to building the tree of life any time soon, even with the rapid accumulation of new complete genome sequences. There is little reason to expect that the rich diversity of species from within the broad clades represented in maximal biclques will soon have many sequences in GenBank. Inclusion of the innumerable “minor” species in the tree of life will therefore require other strategies, such as supertree methods. Nonetheless, biclques will ultimately provide well-supported phylogenetic backbones that can form the basis for more comprehensive supertree-style studies.

An important lingering question concerns the treatment of overlapping biclques. Although some biclques found by running this algorithm are mutually exclusive, many are not. The two “optimal” biclques shown in figure 2, for example, contain many of the same se-

quences, and these overlap as well with some of the other maximal bicliques that satisfy the input constraints. If it were desirable to construct many phylogenetic trees from these data sets, it would be necessary to develop strategies either to avoid redundantly using the same sequence data or to account for that redundant use in later applications. One way to formalize this problem is to (edge-)partition the bipartite graph with the fewest bicliques possible. Deciding whether it is possible to partition a bipartite graph into k bicliques is *NP*-complete (Amilhastre 1999), but minimizing k would guarantee at least that the average number of sequences (edges) per data set (biclique) was maximized. More specific optimality criteria relating to the entire collection of bicliques might be necessary to build robust large phylogenies, however, such as requiring bicliques to be of a minimum size.

Caveats and Limitations

Concatenation of sequences from different genes may not always be a good idea. If different clusters contain different phylogenetic signals, either because of real differences in their evolutionary history, or because of different statistical biases, concatenation may obscure the underlying species tree (Bull et al. 1993). An extensive literature has considered the problem of combinability of data, and statistical tests are widely available (e.g., in PAUP* and other software). However, until very recently these tests have generally dealt with two or three data sets at a time. Scaling up tests to tens or ultimately perhaps even hundreds of genes will present important new challenges. Judging by recent phylogenetic analyses using concatenated genes, the tendency will be to combine data by default, in the hopes that weight of evidence will resolve any conflicts. As genes are sampled from multiple linkage groups and multiple genomes, however, the chances for conflicts between gene trees will rise (Baker and Desalle 1997; Krzywinski, Wilkerson, and Besansky 2001).

Extensions

The problems described in this paper can be easily modified to include weights on either the taxa or the clusters. A natural weight on clusters is the length of the sequences, which should be crudely correlated with robustness (all else being equal). For phylogenetic reconstruction it might be worthwhile to have one gene of 2,500 nucleotides rather than three genes that together only comprise 1,500 nucleotides. At present we treat all clusters as equal. Equal weighting may have its uses if each cluster is considered a potentially independent source of evidence on the species tree. However, in the case of the green plant data, the bicliques were made almost exclusively of genes in a single linkage group, the chloroplast genome. In that case, weighting by number of sites might be instructive. Better still might be weighting by an a posteriori analysis of the phylogenetic signal in the cluster, such as by an average bootstrap score or posterior probability from Bayesian analysis. However, implementing this sort of analysis would be difficult, because the taxa from the cluster that are eventually represented in

the biclique might be in a part of the tree that is poorly supported, even if the remainder of the tree is strongly supported.

Another straightforward extension is to constrain bicliques to contain specific taxa, clusters, or both. Phylogeneticists may wish to obtain large bicliques that include specific taxa of interest; molecular evolutionists may wish to include genes sampled from specific classes of molecules. The algorithm described in this paper can be easily modified to start from a given set of constraints.

It may be important to extend these methods to “quasi-bicliques”—bicliques that are allowed to have some fixed number or fraction of empty elements (an element being an entire sequence missing for some taxon). It should be possible to construct quasi-bicliques larger than the bicliques found here. Recent large multigene studies (Qiu et al. 1999; Soltis, Soltis, and Chase 1999; Murphy et al. 2001) all contain “holes” in their data matrices, and are thus quasi-bicliques. The problem is to assess the trade-off between better sampling in a quasi-biclique and additional noise owing to the missing data (Kearney 2002). Quasi-bicliques might also be used to identify which new sequences should be obtained in the laboratory to permit true bicliques to be constructed. In this way, algorithms can guide phylogenetic experimental design.

Appendix

Exact Algorithm

Given a cluster set C , and natural numbers k and m , ($k, m \geq 1$), find all the bicliques, $K_{k',m'}$, for C , in which $k' \geq k, m' \geq m$. A *cluster* is defined here as a set of taxon names (for the relevant sequences). A *biclique* will be described by the set of clusters it contains, where it is understood that the joint intersection of taxa in these clusters comprises the taxon elements of the biclique. Define two functions that take a biclique as an argument: $\text{Cluster}(b)$, which returns the clusters of the biclique, and $\text{Intersect_Set}(b)$ which returns the taxa in the biclique, which are obtained from the joint intersection of $\text{Cluster}(b)$.

SET Max-Biclique(C, k, m)

```
{
  /* Initialization */
  Delete from  $C$  any cluster that is found in fewer than  $m$ 
  taxa;
  Delete from  $C$  any taxon that is found in fewer than  $k$ 
  clusters;
  Set  $k' = 1$ ; Set  $\text{current\_biclique\_set} = \emptyset$ ;
  FOREACH cluster,  $c$ , in  $C$ 
  {
    Make a biclique,  $b$ , such that  $\text{Cluster}(b) = \{c\}$ , and
     $\text{Intersect\_Set}(b) = c$ ; Add  $b$  to  $\text{current\_biclique\_set}$ ;
  }
  /* Main loop */
  WHILE ( $\text{current\_biclique\_set} \neq \emptyset$ )
  {
    FOREACH biclique,  $b$ , in  $\text{current\_biclique\_set}$ 
```

```

{
  IF |Intersect_Set(b)| < m THEN
    Delete b from current_biclique_set ;
  ELSE
    IF  $k' \geq k$ , THEN
      Add b to the solutions_biclique_set;
    }
  Set  $k' = k' + 1$ ;
  FOREACH biclique, b, in current_biclique_set
  {
    FOREACH cluster, c in (c Cluster (b))
    {
      Let  $b'$  be a biclique such that  $\text{Cluster}(b') = \text{Cluster}(b) \cup \{c\}$ 
      Add  $b'$  to current_biclique_set;
    }
  }
  FOREACH biclique, b, in current_biclique_set
  {
    IF |Cluster(b)| =  $k' - 1$  THEN
      Remove b, from current_biclique_set ;
    }
  }
  If  $k' < k$ , THEN
    RETURN  $\emptyset$ ; /* no bicliques exist for inputs k and m. */
  ELSE
    RETURN solutions_biclique_set
}

```

Acknowledgments

This research was supported by National Science Foundation grant 0075319 to M.J.S. and O.E. We thank two anonymous reviewers for comments.

Literature Cited

- Alexe, G., S. Alexe, Y. Crama, S. Foldes, P. L. Hammer, and B. Simeone. 2002. Consensus algorithms for the generation of all maximal bicliques. DIMACS Technical Report No. 2002-52.
- Altschul, S., W. Gish, W. Miller, E. W. Myers, and D. Lipman. 1990. A basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.
- Amilhastre, J. 1999. Complexity of minimum biclique decomposition of bipartite graphs. Université Montpellier II/CNRS. <http://citeseer.nj.nec.com/148735.html>.
- Baker, R. H., and R. Desalle. 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Syst. Biol.* **46**:654-673.
- Baptiste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Durufle, T. Gaasterland, P. Lopez, M. Muller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc. Natl. Acad. Sci. USA* **99**:1414-1419.
- Bininda-Emonds, O. R. P., S. G. Brady, J. Kim, and M. J. Sanderson. 2001. Scaling of accuracy in extremely large phylogenetic trees. *Pacific Symposium on Biocomputing*. **6**: 547-558.
- Brown, J. R., C. J. Douady, M. J. Italia, W. E. Marshall, and M. J. Stanhope. 2001. Universal trees based on large combined protein sequence data sets. *Nat. Genet.* **28**:281-285.
- Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. *Syst. Biol.* **42**:384-397.
- Chase, M. W., D. E. Soltis, R. G. Olmstead, D. Morgan, D. H. Les, B. D. Mishler, M. R. Duvall, R. A. Price, H. G. Hills, and Y.-L. Qiu. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Ann. Missouri Bot. Gard.* **80**:528-580.
- Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein. 2001. Introduction to algorithms, 2nd edition. MIT Press, Cambridge, Mass.
- Daubin, V., M. Gouy, and G. Perriere. 2001. Bacterial molecular phylogeny using supertree approach. *Genome Informatics* **12**: 155-164.
- Dawande, M., P. Keskinocak, J. Swaminathan, and S. Tayur. 2001. On bipartite and multipartite clique problems. *J. Algorithms* **41**:388-403.
- Dondoshansky, I. 2002. Blastclust (NCBI Software Development Toolkit), 6.1 edition. NCBI, Bethesda, MD.
- Erdos, P. L., M. A. Steel, L. A. Szekely, and T. J. Warnow. 1999. A few logs suffice to build (almost) all trees (I). *Random Struct. Algorithms* **14**:153-184.
- Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**:401-410.
- Garey, M. R., and D. S. Johnson. 1979. Computers and intractability: a guide to the theory of NP-completeness. W. H. Freeman, San Francisco.
- Graham, S. W., and R. G. Olmstead. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *Am. J. Bot.* **87**:1712-1730.
- Hillis, D. M. 1996. Inferring complex phylogenies. *Nature* **383**: 130-131.
- Hochbaum, D. S. 1998. Approximating clique and biclique problems. *J. Algorithms* **29**:174-200.
- Kearney, M. 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. *Syst. Biol.* **51**:369-381.
- Krause, A., J. Stoye, and M. Vingron. 2000. The SYSTEMS protein sequence cluster set. *Nucleic Acids Res.* **28**: 270-272.
- Krivtseva, E. V., W. Fleischmann, E. M. Zdobnov, and R. Apweiler. 2001. CluSTR: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res.* **29**:33-36.
- Krzywinski, J., R. C. Wilkerson, and N. J. Besansky. 2001. Toward understanding Anophelinae (Diptera, Culicidae) phylogeny: insights from nuclear single-copy genes and the weight of evidence. *Syst. Biol.* **50**:540-556.
- Lee, Y., R. Sultana, G. Perrea, J. Cho, S. Karamycheva, J. Tsai, B. Parvizi, F. Cheung, V. Antonescu, J. White, I. Holt, F. Liang, and J. Quackenbush. 2002. Cross-referencing eukaryotic genomes: TIGR orthologous gene alignments (TOGA). *Genome Res.* **12**:493-502.
- Liu, F.-G. R., M. M. Miyamoto, N. P. Freire, P. Q. Ong, M. R. Tennant, T. S. Young, and K. F. Gugel. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* **291**:1786-1789.
- Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* **409**:610-614.
- Miya, M., and M. Nishida. 2000. Use of mitogenomic information in teleostean molecular phylogenetics: a tree-based exploration under the maximum-parsimony optimality criterion. *Mol. Phylog. Evol.* **17**:437-455.
- Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* **409**:614-618.

- Peeters, R. 2000. The maximum edge biclique problem is NP-complete. Preprint.
- Qiu, Y.-L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* **402**:404–407.
- Remm, M., C. E. V. Storm, and E. L. L. Sonnhammer. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314**:1041–1052.
- Sanderson, M. J., A. Purvis, and C. Henze. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends Ecol. Evol.* **13**: 105–109.
- Savolainen, V., M. W. Chase, S. B. Hoot, C. M. Morton, D. E. Soltis, C. Bayer, M. F. Fay, A. Y. de Bruijn, S. Sullivan, and Y.-L. Qiu. 2000. Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. *Syst. Biol.* **49**:306–362.
- Soltis, D. E., P. S. Soltis, M. E. Mort, M. W. Chase, V. Savolainen, S. B. Hoot, and C. M. Morton. 1998. Inferring complex phylogenies using parsimony: an empirical approach using three large DNA data sets for angiosperms. *Syst. Biol.* **47**:32–42.
- Soltis, P. S., D. E. Soltis, and M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* **402**:402–404.
- Swofford, D. L. 2002. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Mass.
- Swofford, D. L., G. J. Olsen, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pp 407–514 in D. M. Hillis, C. Moritz, and B. K. Mable, eds. *Molecular systematics*. Sinauer Associates, Sunderland, Mass.
- Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**:22–28.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.
- West, D. B. 2001. *Introduction to graph theory*, 2nd edition. Prentice-Hall, Upper Saddle River, N.J.
- Wiens, J. J. 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? *Syst. Biol.* **47**: 625–640.
- Wolf, Y. I., I. B. Rogozin, N. V. Grishin, and E. V. Koonin. 2002. Genome trees and the tree of life. *Trends Genet.* **18**: 472–479.

Mark Ragan, Associate Editor

Accepted January 29, 2003