

# Identifying optimal incomplete phylogenetic data sets from sequence databases

Changhui Yan<sup>a,b</sup>, J. Gordon Burleigh<sup>a,c,\*</sup>, Oliver Eulenstein<sup>a</sup>

<sup>a</sup> Department of Computer Science, Iowa State University, Ames, IA 50011, USA

<sup>b</sup> Bioinformatics and Computational Biology Graduate Program, Iowa State University, Ames, IA 50011, USA

<sup>c</sup> Section of Evolution and Ecology, University of California, Davis, CA 95616, USA

Received 20 February 2004; revised 7 September 2004

Available online 21 March 2005

## Abstract

We introduce a new method for identifying optimal incomplete data sets from large sequence databases based on the graph theoretic concept of  $\alpha$ -quasi-bicliques. The quasi-biclique method searches large sequence databases to identify useful phylogenetic data sets with a specified amount of missing data while maintaining the necessary amount of overlap among genes and taxa. The utility of the quasi-biclique method is demonstrated on large simulated sequence databases and on a data set of green plant sequences from GenBank. The quasi-biclique method greatly increases the taxon and gene sampling in the data sets while adding only a limited amount of missing data. Furthermore, under the conditions of the simulation, data sets with a limited amount of missing data often produce topologies nearly as accurate as those built from complete data sets. The quasi-biclique method will be an effective tool for exploiting sequence databases for phylogenetic information and also may help identify critical sequences needed to build large phylogenetic data sets.

© 2005 Elsevier Inc. All rights reserved.

**Keywords:** Maximal biclique; Quasi-biclique; Missing data; Supermatrix

## 1. Introduction

The rapidly growing amount of DNA and protein sequence data provides a wealth of information from which to infer evolutionary histories. Evolutionary biologists are now challenged to find ways to optimally utilize existing sequence data (Sanderson and Driskell, 2003), and today many phylogenetics and comparative biology studies use some sequence data obtained from public databases. With the availability of genomic data, several recent phylogenomic studies have assembled large sequence alignments with data mostly or exclusively obtained from public databases (Bapteste et al., 2002; Driskell et al., 2004; Lerat et al., 2003; Rokas

et al., 2003). Furthermore, constructing large sections of the Tree of Life will doubtlessly require combining sequence data from many different studies (Craclaw and Donoghue, 2004). Yet few formal methods exist for exploiting sequence databases, like GenBank (<http://www.ncbi.nlm.nih.gov>) or Swiss-Prot (<http://us.expasy.org/>), to obtain the best possible phylogenetic data sets (Sanderson et al., 2003). The most useful phylogenetic data set will include many taxa and genes with a limited amount of missing data. Identifying such data sets requires optimizing the tradeoff between increasing gene and taxon sampling and limiting the missing data. We present a new method that uses graph theoretic techniques to identify optimal data sets from large sequence databases.

Sequence databases often have sparse distributions of sequences among all taxa. A few taxa (such as *Arabidopsis thaliana* and *Drosophila melanogaster*) have many

\* Corresponding author. Fax: +1 530 752 1449.

E-mail address: [jgburleigh@ucdavis.edu](mailto:jgburleigh@ucdavis.edu) (J.G. Burleigh).

sequences, and a few genes (such as *rbcL* in plants or *coxI* in animals) have been sequenced from many taxa (Sanderson and Driskell, 2003). However, there are few large sets of taxa that have many common gene sequences (Sanderson and Driskell, 2003; Sanderson et al., 2003), and even data sets from many large phylogenetic studies contain some missing gene sequences (e.g., Murphy et al., 2001; Qiu et al., 1999). Recently, Sanderson et al. (2003) described an approach to identify the largest *complete data sets* from sequence databases. In a complete data set, each taxon has sequence data for all genes. However, it may be possible to identify much larger data sets if the search method allows a limited amount of missing data, or *holes*, in the data set. Though missing data can be problematic for phylogenetic inference, a limited number of holes may have little or no effect (e.g., Kearney, 2002; Wiens, 1998). We adapt the approach of Sanderson et al. (2003) to present a new method for identifying optimal *incomplete data sets*, or data sets that contain some holes, and demonstrate the utility of this method with simulated databases and a set of green plant sequences from GenBank.

## 2. Materials and methods

### 2.1. Definitions

The method  $\alpha$ -quasi-biclique sampling, or quasi-biclique sampling for short, searches for optimal incomplete data sets from sequence databases in a way that limits the amount of missing data and ensures that there is necessary taxonomic overlap among gene sequences. First, a sequence database is represented as a bipartite graph (Fig. 1). A bipartite graph has two sets of nodes, one representing the genes in the database and the other representing the taxa (Fig. 1). An edge connects a taxon node and a gene node if the gene sequence exists for the taxon (Fig. 1). The graph representation of a complete data set is a *biclique* (Fig. 1; Sanderson et al., 2003). Within a biclique, edges connect all gene nodes with all taxon nodes. If a biclique cannot be extended by adding additional genes or taxa, it is a *maximal biclique* (Fig. 1). Thus, the largest complete data sets in a database can be identified using graph theoretic methods that identify maximal bicliques (Sanderson et al., 2003).

Quasi-biclique sampling extends complete data sets (maximal bicliques) by adding taxa that have sequence data from a specified percentage of the genes in the original maximal biclique and/or adding genes for which a specified percentage of the taxa have sequence data (Fig. 2). When a bipartite graph lacks some edges, it is a *quasi-biclique*, and the specified percentage of missing edges is the  $\alpha$ -level of the quasi-biclique. For example, if a maximal biclique contains three genes and three taxa, a 66%-quasi-biclique also would include all the taxa with

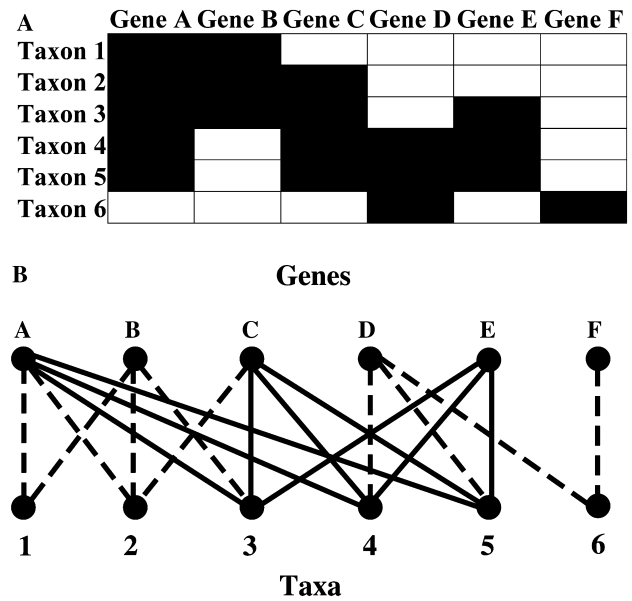


Fig. 1. An example of a biclique representation of a sequence data matrix. (A) Sequence data matrix with six genes and six taxa. The filled (black) cells represent sequence data, and the empty cells represent holes or missing data. (B) The data matrix as a bipartite graph. An edge connects a taxon and a gene if the taxon has sequence data for the gene. The bipartite graph  $\{(A,C,E),(3,4,5)\}$  is a 3 by 3 maximal biclique, and its edges have solid lines.

sequence data from two of the genes in the maximal biclique and/or any genes with sequence data from two of the taxa in the maximal biclique (Fig. 2). The amount of holes in the quasi-biclique data set can be controlled by changing the  $\alpha$ -level. If only taxa or only genes are extended, then the minimum percentage of data present, or *saturation*, in the quasi-biclique data set is  $\alpha\%$  no matter how many taxa or genes are added. However, if both the genes and taxa are extended, there is no minimum saturation. The formal definitions and description of the method follow.

A sequence database  $S$  is represented as the bipartite graph  $G = (X, Y, E)$ . The node sets  $X$  and  $Y$  represent the genes and the taxa of the database sequences, respectively. An edge  $\{x, y\} \in E$  exists if and only if the database contains a sequence for gene  $x$  and taxon  $y$ . A *complete data set* in the sequence database  $S$  is a *biclique* in  $G$ , that is an ordered pair  $(X_B, Y_B)$  such that  $X_B \subseteq X$  and  $Y_B \subseteq Y$  where  $\{x, y\} \in E$  for any  $x \in X_B$  and  $y \in Y_B$ . The biclique  $(X_B, Y_B)$  is *maximal*, if there is no biclique  $(X'_B, Y'_B)$  such that  $\{\{x, y\} : x \in X_B, y \in Y_B\} \subset \{\{x, y\} : x \in X'_B, y \in Y'_B\}$ .

An  $\alpha$ -extension of a biclique  $(X_B, Y_B)$  is an ordered pair  $(X_E, Y_E)$  where  $X_E \subseteq (X - X_B)$  and  $Y_E \subseteq (Y - Y_B)$  such that at least  $\alpha\%$  of the nodes in each of the node sets  $X_B$  and  $Y_B$  are connected through edges to all nodes in  $X_E$  and  $Y_E$ , respectively. An  $\alpha$ -extension  $(X_E, Y_E)$  of a biclique  $(X_B, Y_B)$  is *maximal*, if  $|X'_E| \leq |X_E|$  and  $|Y'_E| \leq |Y_E|$  for any  $\alpha$ -extension  $(X'_E, Y'_E)$  of the biclique  $(X_B, Y_B)$ . Note that any biclique has a maximal  $\alpha$ -exten-

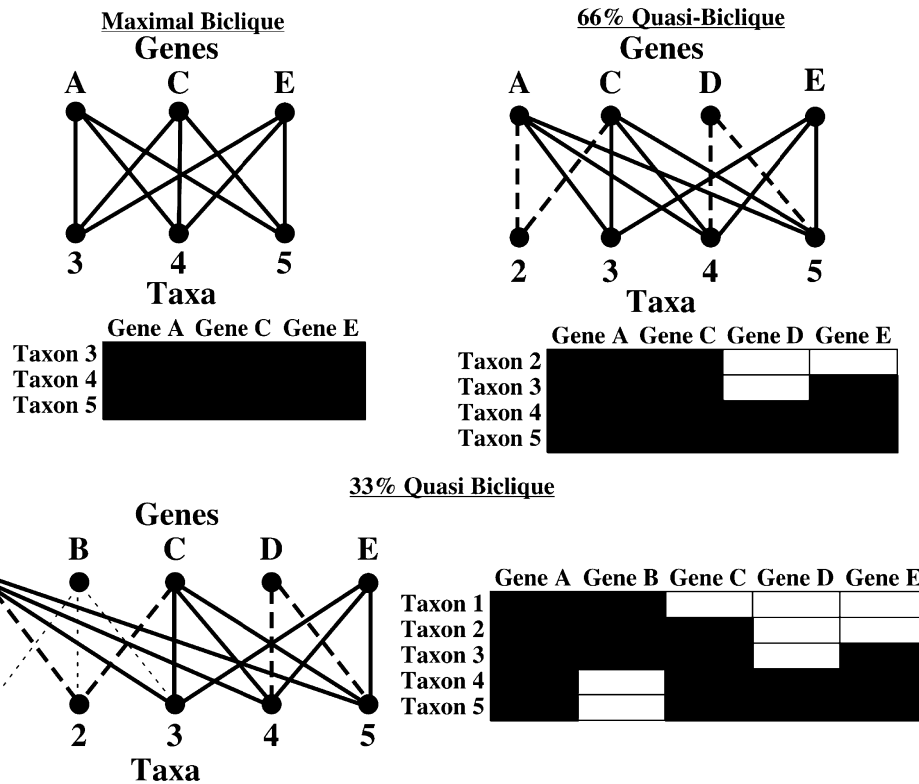


Fig. 2. Examples of the  $\alpha$ -quasi-biclique search strategy of the data set from Fig. 1. First, a maximal biclique,  $\{(A,C,E),(3,4,5)\}$ , is identified, and this represents a complete data set. Next the maximal biclique is extended. The 66%-quasi-biclique extension adds all genes or taxa with edges connected to 2 of the 3 nodes of a node set in the original maximal biclique. The new edges are shown with dashed lines, and the saturation of the 66%-quasi-biclique  $\{(A,C,D,E),(2,3,4,5)\}$  is  $13/16 = 81\%$ . The 33%-quasi-biclique adds all taxa and genes with edges connected to 1 of the 3 nodes in the node set in the original maximal biclique. The new edges are shown with the light dashed lines, and the saturation of the 33%-quasi-biclique data set is  $17/25 = 68\%$ .

sion that is unique. An  $\alpha$ -quasi-biclique is an ordered pair  $(X_B \cup X_E, Y_B \cup Y_E)$  where  $(X_B, Y_B)$  is a maximal biclique and  $(X_E, Y_E)$  its maximal  $\alpha$ -extension. The saturation of a quasi-biclique  $(X_Q, Y_Q)$  is defined as  $s(X_Q, Y_Q) = |\{(x, y) \in E: x \in X_Q \text{ and } y \in Y_Q\}| / (|X_Q| |Y_Q|)$ , that is the number of edges between nodes from  $X_Q$  and  $Y_Q$  normalized by all possible edges between nodes from  $X_Q$  and  $Y_Q$ .

## 2.2. Simulation study

The performance of the quasi-biclique method was examined using 50 large simulated databases. Each simulated database contains 700 taxa. Each taxon has 50,000 base pairs (bp) of sequence data, representing 100 genes that are each 500 bp in length. The 7:1 species to gene ratio is approximately that found in the Swiss-Prot sequence database. To simulate the data, a 700-taxon model tree was generated for each database using the default parameters of the YULE\_C procedure in r8s (Sanderson, 2003). This produced trees according to the conditional Yule birth process that fixes the time between the root of the tree and the tips (Ross, 2000). Sequences were generated for each model tree using Seq-Gen (Rambaut and Grassly, 1997). The sequences were simulated

according to the Kimura 2-parameter (K2P; Kimura, 1980) model, with equal nucleotide frequencies and a transition to transversion ratio of 2:1. Among site rate variation also was incorporated into the simulations using with a four-category, discrete gamma distribution with a shape parameter,  $\alpha$ , of 0.5 (Yang, 1994). Since all sites in the simulated alignment are independent and identically distributed, each consecutive 500 bp block was considered a separate gene. A consecutive sequence of 500 bp in a block for single taxon is called a cell. The complete simulated database with no missing information is a complete database. To create a sparse distribution of sequences in the simulated databases, in 10 sequence databases, cells were deleted by assigning each cell an 80% chance of deletion, and in 40 simulated databases, cells were assigned a 90% chance of deletion. After the cells are deleted, the simulated matrix is an  $x\%$  incomplete database, where  $x$  is the percentage of deleted cells.

## 2.3. Quasi-biclique search

Maximal bicliques were identified in the incomplete databases using the algorithm of Alexe et al. (2002; see Sanderson et al., 2003). Maximal bicliques with at least five genes and five taxa were identified in the 80% incom-

plete databases, and maximal bicliques with at least three genes and five taxa were identified in the 90% incomplete databases. Each maximal biclique was extended to a  $\alpha$ -quasi-biclique using all possible values of  $\alpha$ . For example, in a five by five maximal biclique, the  $\alpha$  level can be 80, 60, 40, or 20%.

The effect of extending the maximal bicliques on phylogenetic accuracy was examined by comparing trees built from the quasi-bicliques with trees built from complete data sets of the same genes and taxa. Phylogenetic inference was done using maximum parsimony (MP) with PAUP\* (Swofford, 2003). The parsimony heuristic searches uses a TBR branch swapping starting from a tree constructed with random sequential addition, and a maximum of five parsimonious trees were saved. If the heuristic search found more than one equally parsimonious tree, a majority rule consensus tree of the optimal trees was constructed. This heuristic search strategy may not be considered thorough for the larger data sets. However, the purpose of this study is to compare trees made from incomplete and complete data sets, and since the heuristic was applied to both data sets, it should not affect this comparison.

The accuracy of each parsimony tree was measured by calculating the maximum agreement subtree (MAST; Gordon, 1980; Kubicka et al., 1992) of the MP tree and the true model tree using PAUP\* (Swofford, 2003). The *MAST score* is the number of leaves (taxa) in the MAST divided by the total number of leaves in the MP tree (e.g., Chen et al., 2003; Eulenstein et al., 2004). If the MP tree and the true model tree have identical topologies for their shared taxa, then the MAST score would be 1, and if the MP tree differs from the true topology, the MAST score will be less than one. To estimate the effect of the missing cells on the phylogenetic inference, the MAST scores from the MP trees constructed from the quasi-biclique data sets were compared to the MAST scores from MP trees made from the corresponding complete data sets. The *normalized MAST score* is the MAST scores from the quasi-biclique MP tree divided by the MAST score from the complete data set MP tree. If the two trees are identical, the normalized MAST score will be one. If the quasi-biclique MP tree is more accurate (closer to the true model tree) than the complete data set MP tree, then the normalized MAST will be greater than one, and if the complete data set MP tree is more accurate, then the normalized MAST will be less than one.

#### 2.4. Empirical data set

The distribution of missing data in the simulated data sets likely differs from real data sets. Therefore, we also tested the quasi-biclique search method on a real data set from GenBank. Sanderson et al. (2003) used a cluster set of green plant genes from GenBank

to identify the largest complete data sets. The sequence data was extracted from the GBPLN flat files in release 127.0 of GenBank from December, 2001. Sanderson et al. (2003) identified 657 “clusters,” or putative orthologous genes with sequences from at least four taxa, in the original GenBank data. Overall, the clusters contain at least one sequence from 10,141 total taxa. There were 14 maximal bicliques with at least 13 taxa and 13 genes in the cluster set. These maximal bicliques are not independent because many of the maximal bicliques share common sequences. Quasi-biclique searches started from each of these 14 maximal bicliques, and the  $\alpha$ -quasi-bicliques were extended from 90 to 10%, in increments of 10%. A sample file describing the distribution of genes in this data set and a script that will identify the quasi-bicliques is available at <http://ginger.ucdavis.edu>.

### 3. Results

#### 3.1. Simulation study

##### 3.1.1. Identifying quasi-bicliques

On average, 10 maximal bicliques with at least five taxa and five genes were found in each 80% incomplete database, and 15 maximal bicliques with at least five taxa and three genes were found in the 90% incomplete database. The number of taxa rapidly increased as the maximal bicliques was extended into quasi-bicliques, though the increase was slower in the 90% than the 80% incomplete databases (Table 1). For example, in

Table 1  
Summary of the  $\alpha$ -quasi-bicliques from simulated data sets

Quasi-biclique (%)	Taxa	Genes	Saturation	Norm. MAST
<i>(A) 80% incomplete database</i>				
80	10.86	5.08	0.91	0.97
60	44.45	6.84	0.58	0.84
50	44.45	13.84	0.41	0.78
40	187.06	14.24	0.30	0.61
30	187.68	38.00	0.24	0.66
20	474.31	38.40	0.22	0.68
<i>(B) 90% incomplete database</i>				
80	6.00	4.00	0.96	1.01
70	7.00	4.00	0.93	1.06
60	24.86	4.13	0.74	0.93
50	24.68	5.05	0.54	0.87
40	26.31	6.05	0.49	0.88
30	193.64	13.46	0.18	0.37
20	194.93	17.75	0.16	0.38

The number in the “Quasi-biclique” column denotes the  $\alpha$  level. The next two columns show the average number of genes and taxa in the  $\alpha$ -quasi-bicliques. The “Norm. MAST” is the ratio of the MAST scores for the phylogeny made with missing data to the phylogeny made without any missing data (from the complete database). (A) The  $\alpha$ -quasi-bicliques found in the 80% incomplete database, and (B) The  $\alpha$ -quasi-bicliques from the 90% incomplete database.

the 80% incomplete database, the 20%-quasi-bicliques contain 473 taxa of the 700 taxa on average while in the 90% incomplete databases, they contain 194 taxa on average (Table 1). The number of genes increased to an average of 38 in the 80% incomplete database and 18 in the 90% incomplete database (Table 1).

### 3.1.2. Phylogenetic inference of quasi-biclique data sets

Phylogenies built from quasi-biclique data sets with relatively high saturation are often at least as accurate as phylogenies built from complete data sets (Table 1; Fig. 3). The average normalized MAST score for the 80%-quasi-bicliques is close to 1 in both the 80 and 90% incomplete databases (Table 1). In the 80% incomplete databases, trees made from 60%-quasi-bicliques, which have on average 41% of their cells missing, still have a MAST score that is 84% as high as the MAST score of trees made without any missing data (Table 1A, Fig. 3A). In other words, 41% missing

data results in only a 16% loss of accuracy. In the 80% incomplete database, the average normalized MAST scores never decrease below 60%, even when more than 75% of the cells are missing (Table 1A). In the 90% incomplete databases, trees made from 40%-quasi-bicliques (with 49% saturation) are nearly 90% as accurate as trees made without holes (Table 1B). In the 90% incomplete databases, the number of taxa and genes rapidly increases at the 30%  $\alpha$  level, and at this point, the normalized MAST scores decrease markedly (Table 1B).

### 3.2. Green plant database

The  $\alpha$ -quasi-biclique search of the green plant cluster data sets added many genes to the data sets while adding very few holes (Fig. 4A; Table 2). The 80%-quasi-bicliques have 14 genes more than the maximal bicliques on average, and the average saturation is still greater

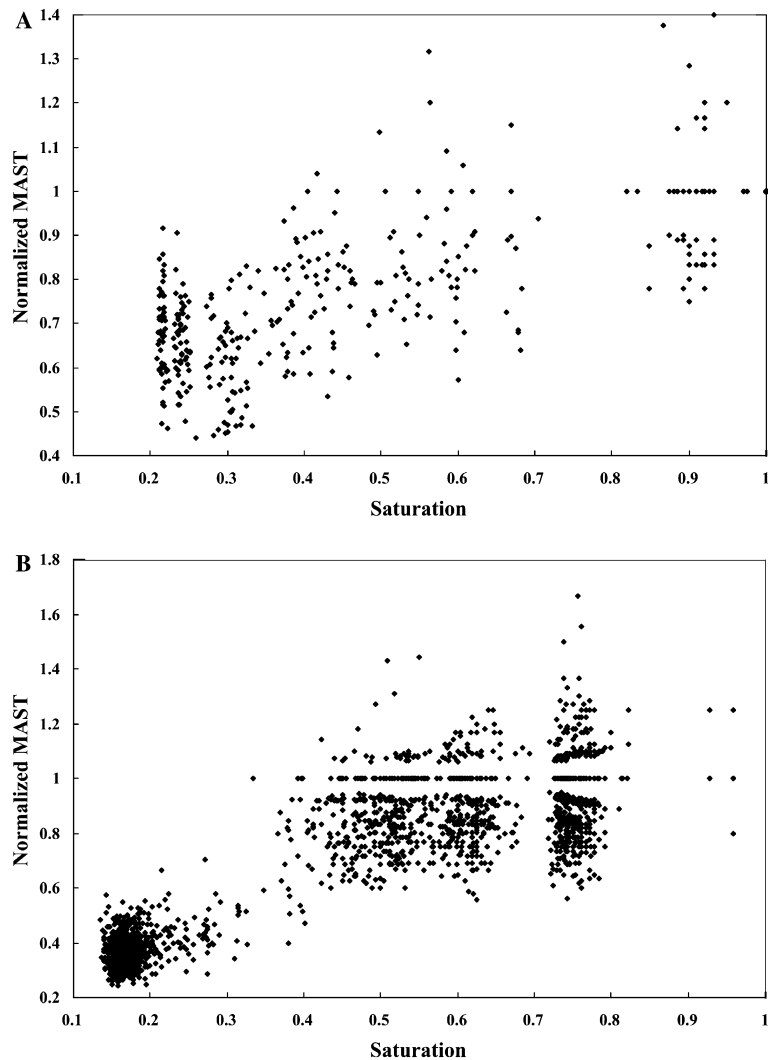


Fig. 3. Graph of  $\alpha$ -quasi-biclique saturation versus the normalized MAST score. Each point on the graph represents a different quasi-biclique. (A) was created from the 80% incomplete database, and (B) was created from the 90% incomplete simulated databases.

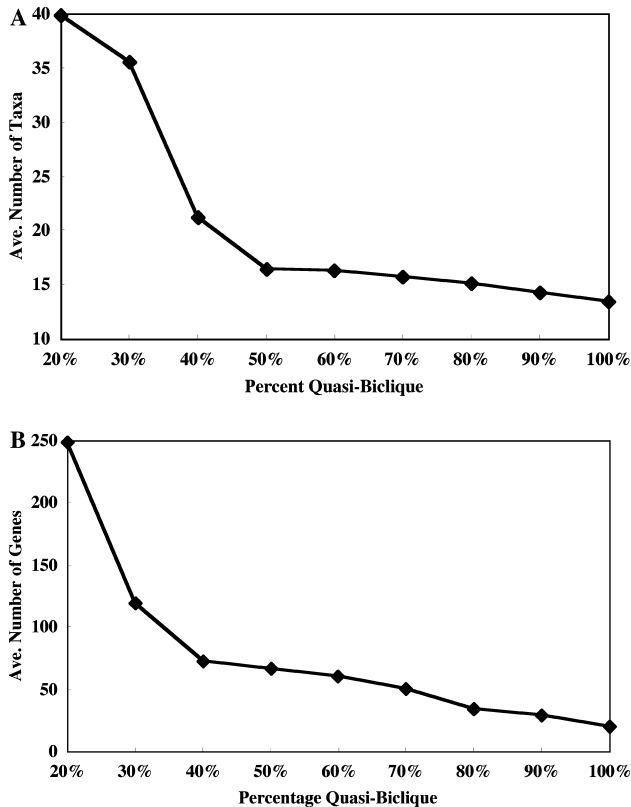


Fig. 4. Graph of the average size of  $\alpha$ -quasi-biclques made from a data set of green plant genes from GenBank. Quasi-biclque data sets were constructed from 14 maximal biclques that contain at least 13 genes and 13 taxa found in set of putative orthologous genes from green plants that were extracted from GenBank. The line shows the change in the average number of taxa (A) or genes (B) in the data sets as the  $\alpha$  level changed.

Table 2  
Average  $\alpha$ -quasi-biclques from a set of green plant sequences from GenBank

Quasi-biclque (%)	Taxa	Genes	Saturation
100	13.5	20.4	1.000
90	14.3	29.6	0.964
80	15.1	34.5	0.930
70	15.7	50.6	0.842
60	16.3	60.4	0.792
50	16.7	67.0	0.752
40	21.3	73.1	0.611
30	35.6	119.2	0.301
20	39.9	248.5	0.181

There were 14 maximal biclques with at least 13 taxa and 13 genes in a set of 657 clusters of putative orthologs in green plant taxa found in GenBank. A quasi-biclque search was performed starting from these maximal biclques using  $\alpha$  levels from 20 to 90%, in 10% increments. Table shows the average number of genes and taxa, as well as the average saturation, in the  $\alpha$ -quasi-biclques. The number in the "Quasi-biclque" column denotes the  $\alpha$  level.

than 90% (Table 2). In the 70%-quasi-biclque searches, the average number of genes has more than doubles to over 50 while the average saturation is still 84% (Table

2). The number of genes increases very rapidly at the 30 and 20%  $\alpha$  level (Fig. 4A, Table 2). The number of taxa in each  $\alpha$ -quasi-biclque increases little until the 20 or 30%  $\alpha$  level (Fig. 4B, Table 2). The 20%-quasi-biclque search added 26 taxa on average to each data set (Table 2). The overall saturation level of the  $\alpha$ -quasi-biclques remains relatively high, at least compared to the  $\alpha$  level, until the 30%  $\alpha$  level (Table 2).

#### 4. Discussion

The  $\alpha$ -quasi-biclque search method can greatly increase the taxon and gene sampling of complete data sets while adding only a small percentage of holes. In the simulated 90% incomplete databases, over four times as many taxa were in 60%-quasi-biclque data sets than the 80%-quasi-biclque data sets, and in the set of green plant genes from GenBank, the average number of genes per data set increased from 20 to 50 while still having only 15% missing data (Table 2). The distribution of the sequences in the database determines the performance of the quasi-biclque search. In the simulated data sets, taxon sampling rapidly increases, but the number of genes increases little until the 50%-quasi-biclques (Table 1). This likely is due to the presence of seven times as many taxa as genes and the random pattern of missing data. In contrast, the number of green plant genes increases rapidly, while the number of taxa remains low (Fig. 4, Table 2). In the GenBank data set, most taxa have very few sequences. The taxon sampling could be increased by reducing the number of genes in the original maximal biclque or by reducing  $\alpha$  even further.

The distribution of genes among taxa in the simulated data sets likely is different from that found in most empirical data sets. Each sequence database will have a different distribution of sequences among taxa, and the distribution of sequences may vary among taxonomic groups and change through time. Thus, it is difficult to determine what a realistic pattern of missing data would be. However, since the  $\alpha$ -quasi-biclque search method greatly increased sampling in databases that have very different patterns of missing data, it likely will be useful in many different databases.

Though quasi-biclque data sets often are much larger than complete, maximal biclque data set, they will be of limited use for phylogenetics if the holes in the matrices greatly reduce phylogenetic accuracy. The simulation experiment suggests that a limited amount of missing data may have little, if any, effect on the accuracy of phylogenetic inference. In the 80% incomplete databases, the 60%-quasi-biclques had 42% holes on average, but the normalized MAST was 84%. Thus, even though the 42% of the data is missing, their accuracy is only 16% less than it would be without missing data

(Table 1). In the 90% incomplete databases, the 40%-quasi-biclques are missing over half of their data, but still have a normalized MAST score of 88% (Table 1). These results indicate that phylogenetic accuracy of large data sets may be remarkably robust to large amounts of missing data, and in some cases, there may be little benefit to filling in all the holes. The effect of missing data on phylogenetic inference is a contentious issue in systematics (e.g., Kearney, 2002; Wiens, 1998, 2003; Wilkinson, 1995). Yet, the results from the simulated data sets are consistent with previous studies suggesting that a limited amount of missing data often has little or no effect on the accuracy of the phylogenetic inference (Kearney, 2002; Wiens, 1998, 2003). Yet, this study was not designed to explicitly test the effect of missing data on phylogenetic inference. For example, unlike many previous simulation studies examining the effect of missing data, both the taxon sampling and the length of the alignment change as the amount of missing data changes. The change in the phylogenetic accuracy may be affected by not only additional missing data but also by changes in the amount of data present and taxon sampling.

The amount of data that may be absent in a data matrix without affecting phylogenetic accuracy likely depends on many factors such as the rate of evolution of the genes, the taxonomic sampling, the shape of the tree, and the method of phylogenetic inference. The percentage of missing data also may not be as important as the amount of data present (Wiens, 2003). In other words, if there are enough informative characters present in the data matrix, then the percentage of missing characters may not matter. For example, in the 80% incomplete database, trees made from the 30% biclques have a slightly greater normalized MAST score than trees made from the 40%-quasi-biclques (Table 1A). The 40 and 30%-quasi-biclques have almost identical taxon sampling, but the 30%-quasi-biclques have 20 more genes on average in the alignment. Thus, even though the 30%-quasi-biclques have lower saturation than the 40%-quasi-biclques, there may be more informative data present in the 30%-quasi-biclque data sets. Similarly, though trees made from 30%-quasi-biclques from the 80% incomplete databases have similar taxon sampling and saturation as 30%-quasi-biclque trees from the 90% incomplete databases, the trees from the 80% incomplete database contain many more genes and also have much higher normalized MAST scores (Table 1). It is difficult to make an a priori recommendation about the acceptable amount of missing data, but as the data sets get larger, the effect of missing data may be ameliorated.

The results of the simulation experiment are also consistent with recent phylogenomic studies using large data matrices with much missing data. While the genes in the simulated database had an equal length and rate

and pattern of evolution, in real data sets, there is much variation in the amount of phylogenetic information among genes as well as in the phylogenetic signal itself (Driskell et al., 2004; Rokas et al., 2003). Still, a recent study of eukaryote phylogeny incorporating 129 genes and roughly 25% missing data indicates that the missing data has little effect on the phylogenetic inference (Philippe et al., 2004). Also, a phylogenetic analysis of a large metazoan data set using sequences obtained from SwissProt demonstrated a remarkably strong phylogenetic signal even with over 90% of the gene sequences missing (Driskell et al., 2004). Thus, it appears that empirical evidence also suggests, it is not necessary to utilize complete data sets and that the much larger incomplete data sets will be useful for phylogenetic inference.

The quasi-biclque method appears to effectively identify optimal phylogenetic data sets from large and sparsely distributed sequence databases, and demonstrates how methods developed from graph theory may be useful for phylogenetic data mining. The quasi-biclque method also may be adapted or improved to expand its utility for phylogenetics. The edges of the bipartite graph may be given different weights to increase the probability that specific genes or taxa are sampled in phylogenetic databases. The quasi-biclque search also may be performed directly on nucleotide data, using taxon by nucleotide graph instead of a taxon by gene graph, to find optimal alignments. Furthermore, searching taxon by tree graphs may be useful for finding optimal data sets for supertree construction. Driskell et al. (2004) propose a different method for building incomplete data sets that involves concatenating complete data sets that have a specified amount of taxonomic overlap, and Bininda-Emonds (2004) suggests a similar approach for building supertrees by combining trees made from complete data sets. Still, the sparseness of the databases remains a problem, and useful phylogenetic data sets may not exist to answer many phylogenetic questions. Such quasi-biclque methods may be useful for identifying identify the critical missing data that need to be filled in order to obtain a sufficiently complete data matrix (Sanderson et al., 2003).

## Acknowledgments

We thank Mike Sanderson and Cecile Ané for helpful discussion and comments on this manuscript. Wen-Chieh Chang assisted implementing the maximal biclque algorithm, and Lisa Thurston provided computer assistance. This manuscript also was improved by the comments of two anonymous reviewers. This work was supported by NSF Grants 1053164 (Burleigh and Eulenstein) and EF-0334832 (Eulenstein).

## References

- Alexe, G., Alexe, S., Crama, Y., Foldes, S., Hammer, P.L., Simeone, B., 2002. Consensus algorithms for the generation of all maximal bicliques. DIMACS Technical Report No. 2002-52.
- Baptiste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Duruffe, L., Gaasterland, T., Lopez, P., Muller, M., Philippe, H., 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., 2004. The evolution of supertrees. Trends Ecol. Evol. 19, 315–322.
- Chen, D., Diao, L., Eulenstein, O., Fernandez-Baca, D., Sanderson, M.J., 2003. Flipping: a supertree construction method. In: DIMACS Series in Discrete Mathematics and Theoretical Computer Science. AMS, Providence, RI, pp. 135–160.
- Cracraft, J., Donoghue, M.J., 2004. Assembling the tree of life: where we stand at the beginning of the 21st century. In: Cracraft, J., Donoghue, M.J. (Eds.), Assembling the Tree of Life. Oxford University Press, New York, pp. 553–561.
- Driskell, A.C., Ané, C., Burleigh, J.G., McMahon, M.M., O'Meara, B.C., Sanderson, M.J., 2004. Prospects for building the tree of life from large sequence databases. Science 306, 1172–1174.
- Eulenstein, O., Chen, D., Burleigh, J.G., Fernandez-Baca, D., Sanderson, M.J., 2004. Performance of flip-supertree construction with a heuristic algorithm. Syst. Biol. 53, 299–308.
- Gordon, A.D., 1980. On the assessment and comparison of classifications. In: Tomassine, R. (Ed.), Analyse De Données Et Informatique. Le Chesnay, INRIA, France, pp. 149–160.
- Kearney, M., 2002. Fragmentary taxa, missing data and ambiguity: mistaken assumptions and conclusions. Syst. Biol. 51, 369–381.
- Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.
- Kubicka, E., Kubicki, G., McMorris, F.R., 1992. On agreement subtrees of two binary trees. Congressus Numerantium 88, 217–224.
- Lerat, E., Daubin, V., Moran, N.A., 2003. From gene trees to organismal phylogeny in prokaryotes: the case of the  $\gamma$ -proteobacteria. PLoS Biol. 1, 1–9.
- Murphy, W.J., Eizirik, E., Johnson, W.E., Zhang, Y.P., Ryder, O.A., O'Brien, S.J., 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409, 614–618.
- Philippe, H., Snell, E.A., Baptiste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21, 1740–1752.
- Qiu, Y.-L., Lee, J., Bernasconi-Quadroni, F., Soltis, D.E., Soltis, P.S., Zanis, M., Zimmer, E.A., Chen, Z., Savolainen, V., Chase, M.W., 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402, 404–407.
- Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.
- Rokas, A., Williams, B.L., King, N., Carroll, S., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.
- Ross, S., 2000. Introduction to Probability Models, seventh ed. Harcourt Academic Press, New York.
- Sanderson, M.J., 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19, 301–302.
- Sanderson, M.J., Driskell, A.C., 2003. The challenge of constructing large phylogenetic trees. Trends Pl. Sci. 8, 374–379.
- Sanderson, M.J., Driskell, A.C., Ree, R.H., Eulenstein, O., Langley, S., 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol. Biol. Evol. 20, 1036–1042.
- Swofford, D.L., 2003. PAUP\*. Phylogenetic analysis using parsimony (\* and other methods). Version 10. Sinauer Associates, Sunderland, MA.
- Wiens, J.J., 1998. Does adding characters with missing data increase or decrease phylogenetics accuracy? Syst. Biol. 47, 625–640.
- Wiens, J.J., 2003. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538.
- Wilkinson, M., 1995. Coping with abundant missing entries in phylogenetic inference using parsimony. Syst. Biol. 44, 548–558.
- Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.