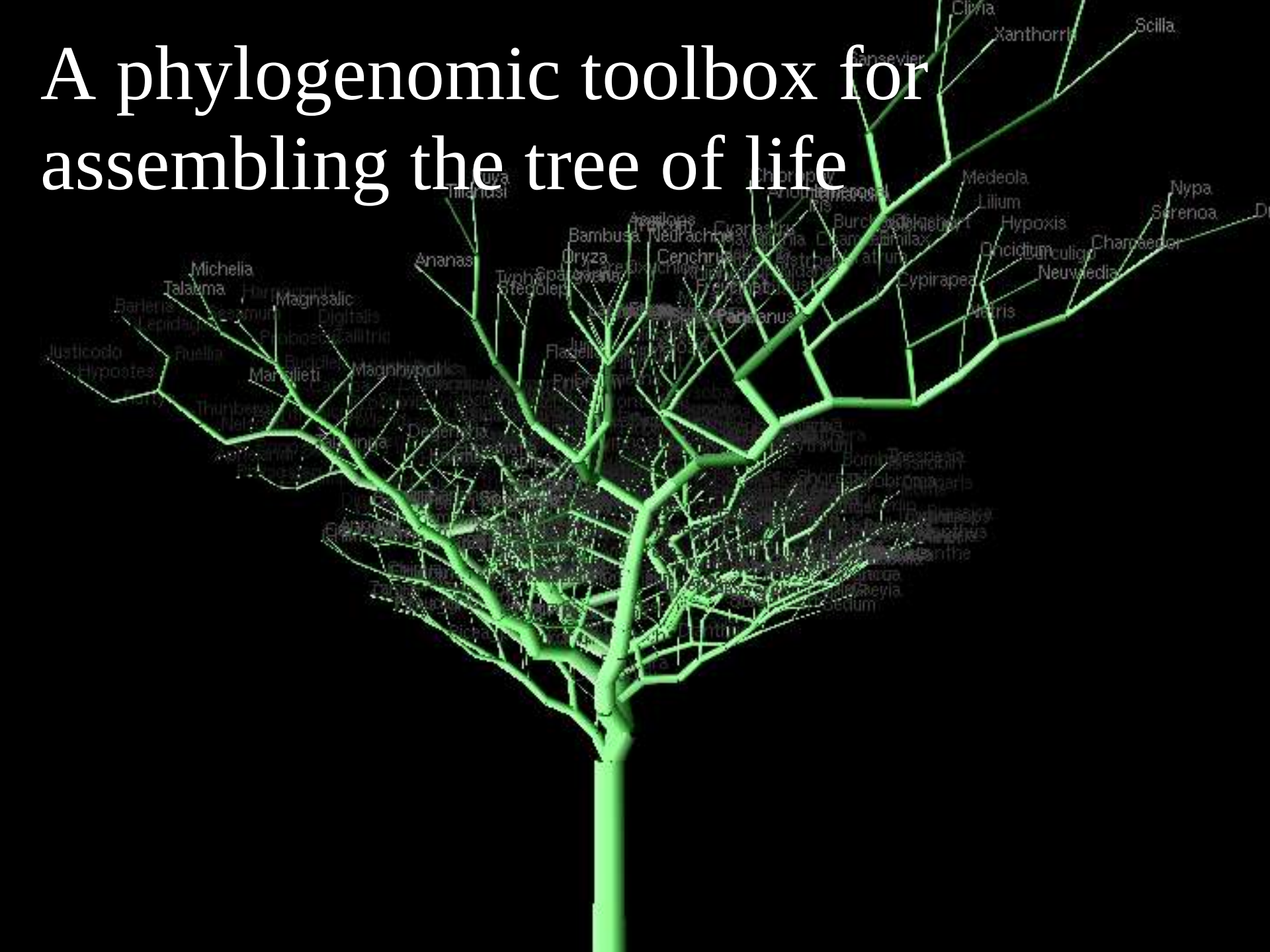
A phylogenomic toolbox for assembling the tree of life



or, The *Phylota* Project (http://www.phylota.org)

UC Davis
Mike Sanderson
Amy Driskell

U Pennsylvania
Junhyong Kim

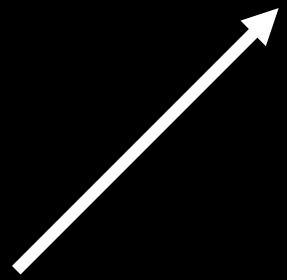
Iowa State
Oliver Eulenstein
David Fernández-Baca

NSF AToL grant DEB-0334963

Overall goal: Developing algorithms and software tools for exploiting the phylogenetic information in large sequence databases

Proteins

Taxa



Increasing density

Sparse data availability matrix for GenBank (rel. 137) green plant proteins (single copy potentially informative)

Density = 0.002

Specific goals

- Characterizing the phylogenetic information content of sequence databases
- Extracting optimal collections of data sets from databases to build species trees
- Designing inputs for supermatrix and supertree construction
- Algorithms for targeted sequence efforts
- Further work on novel supertree methods (Flip-distance; additive distance methods)

Project status

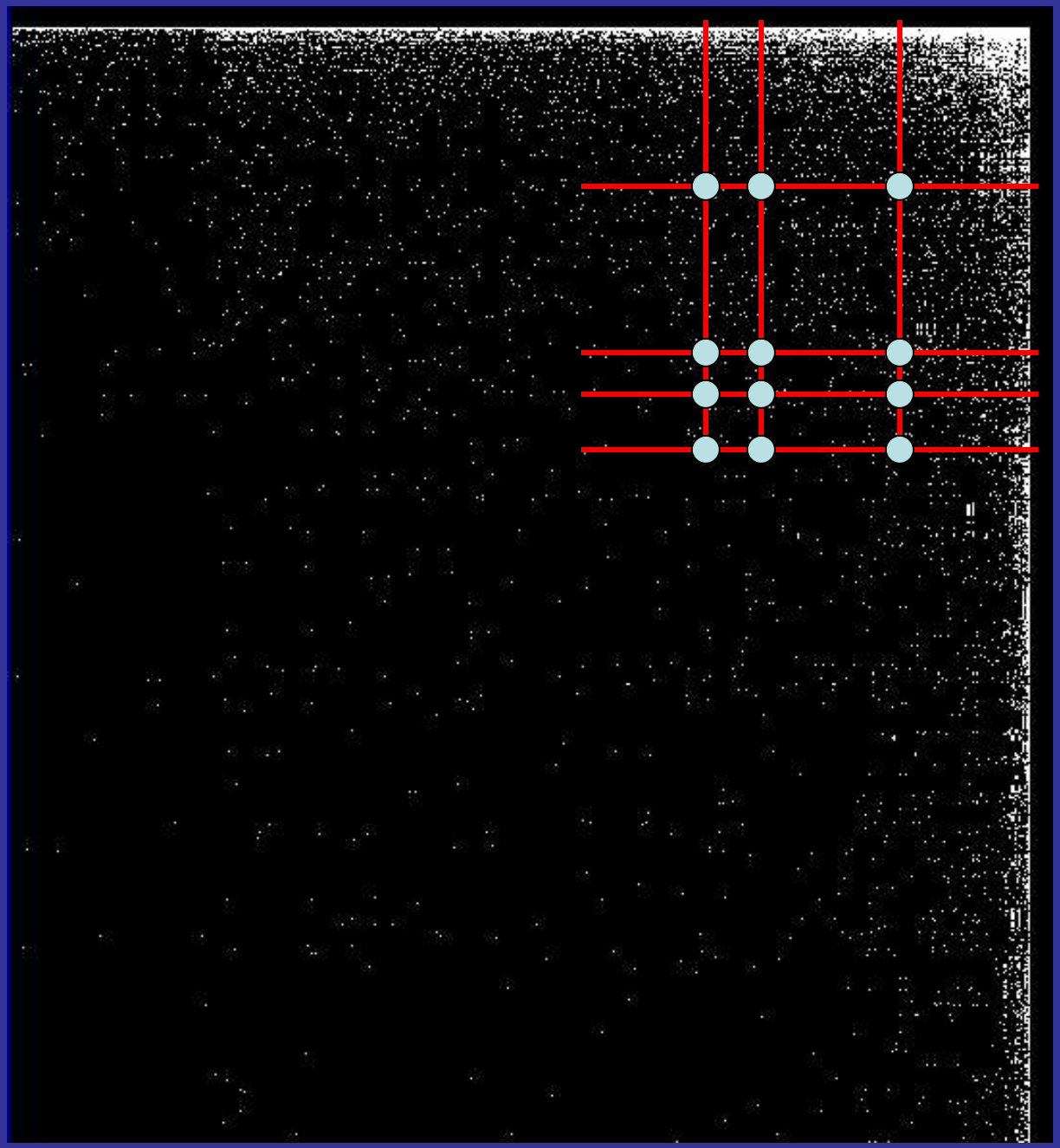
- Year 1 completed
- Benchmark data sets of ~300,000 proteins constructed and posted on web site
- Progress on formal problem definitions for groves, quasi-bicliques, targeted sequencing
- Early prototypes of software tools posted on web site (bicliques, groves, clustering, sample visualization)
- Preliminary experiments with large sparse supermatrices and their associated supertrees

Phylogenetic information content of two sets of protein sequences from large databases (Driskell et al. 2004)

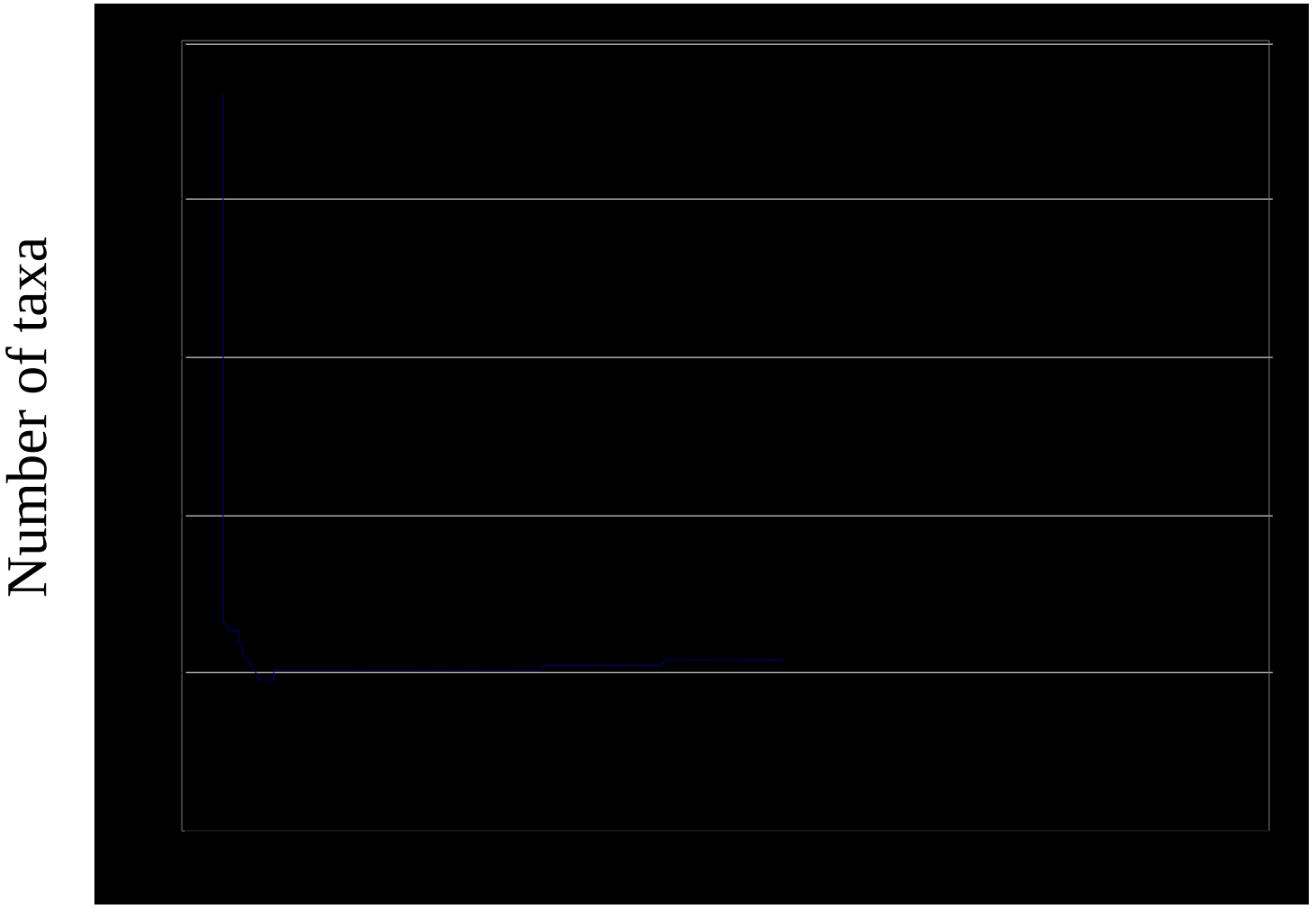
	Swiss-Prot release 40.29	GenBank release 137 (green plant)
Number of sequences in release	121,218	185,418
Total number of sequences clustered	121,218	185,089
Total number of taxa	7449	16,348
Total number of clusters	64,712	59,144
MINIMAL PHYLOGENETIC CLUSTERS		
Number of clusters	4214 (6.5%)	1365 (2.3%)
Sequence coverage	41,812 (34%)	65,113 (35%)
Taxon coverage	5538 (74%)	15,599 (95%)
SINGLE-COPY CLUSTERS		
Number of clusters	3592 (5.6%)	853 (1.4%)
Sequence coverage	28,742 (24%)	39,443 (21%)
Taxon coverage	4404 (59%)	14,502 (89%)
Density	0.0018	0.0021
GROVES		
Minimum number of groves	123	15
Number of orphan clusters	67	7
Minimum number of clusters in largest grove	3183	814
Minimum number of sequences in largest grove	25,272 (21%)	38,700 (49%)
Minimum number of taxa in largest grove	2695 (36%)	14,169 (87%)
BICLIQUES		
Number of nontrivial maximal bicliques	43,576	5587
Sequence coverage	23,855 (20%)	15,092 (8.2%)
Taxon coverage	1449 (19%)	4230 (26%)
Number of clusters in biclique set	3187 (4.9%)	645 (1.1%)
Largest biclique (in terms of taxa)	2×76	2×975
Largest biclique (in terms of sequences)	352×4	70×4

Finding complete supermatrices in a sparse database…

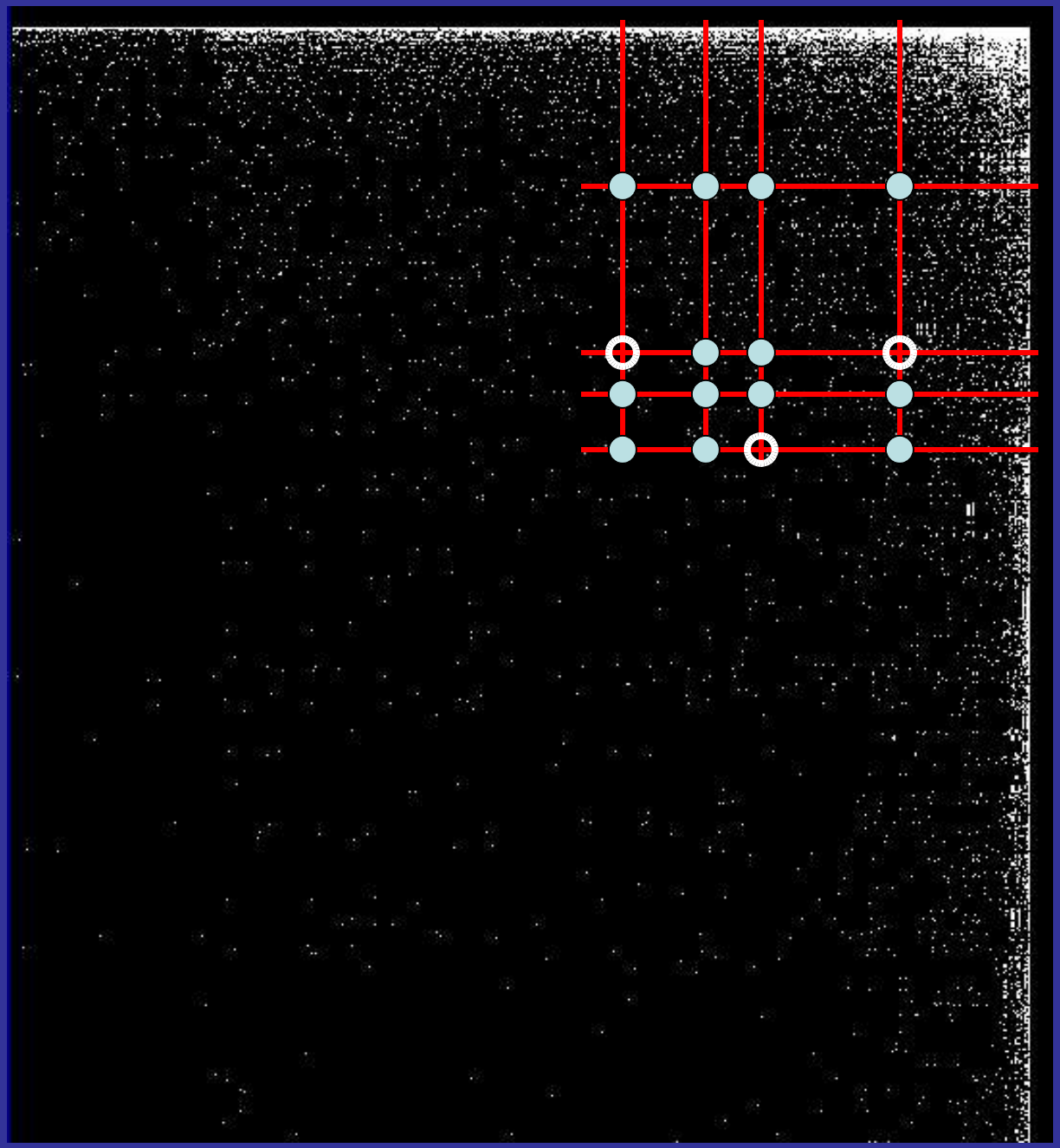
A “complete” matrix is one with no missing proteins



Sizes of the 8300 maximal biclique data sets



Constructing *incomplete* supermatrices in a sparse database. What are efficient methods for assembling these matrices and for identifying optimal



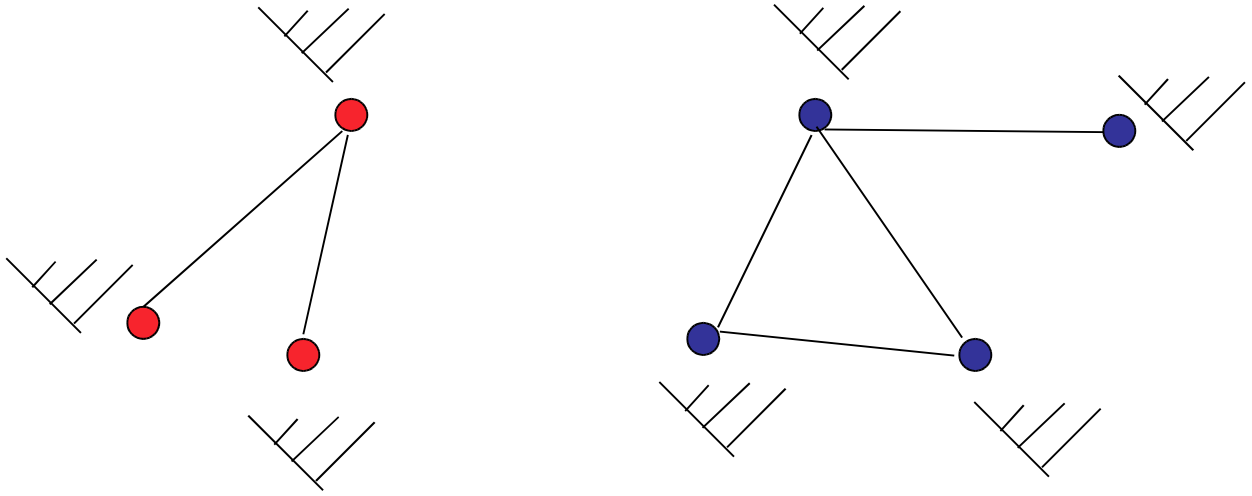
Green plant supermatrix: 69 taxa × 254 proteins

Density = 16% (13% of characters)

96,584 amino acids

2777 sequences

Groves: collections of trees that can be useful inputs to supertree construction (based on taxonomic overlap

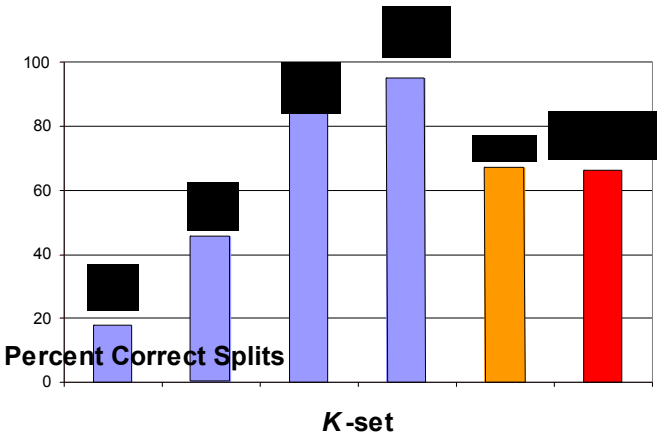


Constructing Super-trees and Consensus trees from distance embedding

Goal: Develop new algorithmic approaches to super-tree and consensus tree construction by characterizing trees as Additive Distance (AD) matrices.

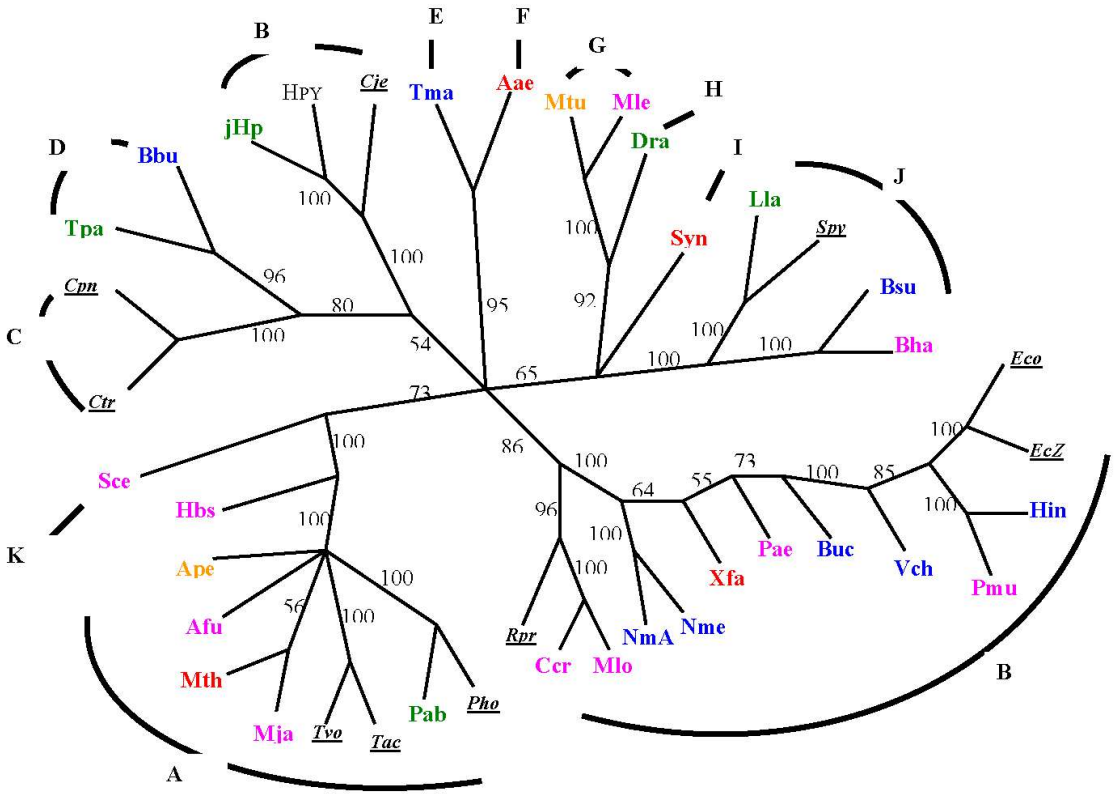
Rationale: For every positive edge-weighted tree, there is an invertible correspondence to an AD matrix. Therefore, we can view construction of super-trees in terms of space of additive distance matrices. Each tree or a gene cluster yields a projection from the hypothetical full distance matrix (corresponding to the full supertree). The problem is to extrapolate to the full matrix (Fig X).

Advantages: For n total input taxa, the space of additive distances has $O(n^2)$ embedding dimensions as opposed to $O(2^n)$ embedding dimensions in usual combinatorial algorithms such as Matrix Representation Parsimony (MRP). The histogram on the right shows a small simulation test where a 16-taxon tree was estimated from a random collection of k quartets. The results suggest that the AD embedding method more efficiently reconstructs the supertree compared to MRP. Furthermore, since the embedding is in metric space, it allows us to use standard methods to incorporate different confidence in branches as well as different confidence in data sets.



The figure on the right shows one result from AD embedding method where we assembled 476 gene clusters of varying sizes and constructed a single supertree (numbers indicate bootstrap values).

Next Steps: The method has been developed for generally dense coverage of submatrices as in the tree shown here. However, for large-scale datasets as found in the Genbank, the submatrices very sparsely cover the relationship of the taxa. Thus many of the missing relationships need to be extrapolated. We are investigating the usage of AD matrix extrapolation method proposed by Makarenkov and Lapointe (Bioinformatics 20:2113).



The tree is for 40 completely sequence microbial genomes with Yeast (Sce). The three letter taxon code can be found at http://ncbi.nlm.nih.gov

Faster Flipping for Supertree Construction

- Improvement on previous flipping program (Chen et al., *Systematic Biology*, 53(2):299–308, 2004)
- New method achieves speedup of at least *n* (= number of taxa) for NNI, SPR, and TBR by avoiding re-computation of flip scores for subtrees

Targeted sequencing algorithms: collaborations with other AToLs

- NemAToL
- Early Bird
- The green tree of life
- Eu-Tree