Supertree Analysis of the Plant Family Fabaceae

Tiffany Morris

Advisor:
Martin Wojciechowski

June 2004-December 2004

Project Goal

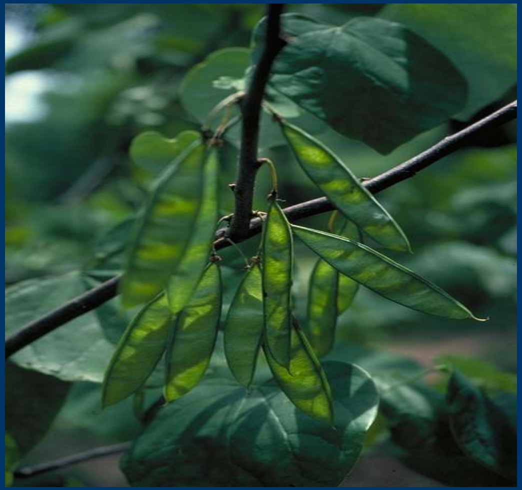
- To obtain a Supertree for the plant family Fabaceae utilizing phylogenetic trees found in previously published studies


Tree of Life

National and international project to collect information on the origin, evolution, and diversity of organisms with the goal of producing a tree of all life on Earth

Fabaceae Family (Legumes)

- Large family of flowering plants
  - 750 genera
  - 18,000 species
  - 3rd largest family, cosmopolitan in distribution
  - Many of these species are agriculturally and economically important
    - Pisum sativum (pea)
    - Medicago sativa (alfalfa)
    - Lens culinaris (lentil)
    - Arachis hypogaea (peanut)
    - Parkinsonia aculeata (palo verde)



Given the basic difficulties with inferring trees of a relative few taxa, how do we infer BIG phylogenies,

with hundreds or thousands of taxa. . .?

The Tree of Life?

Two basic philosophical approaches:

"**total evidence"** approach requires combined data to be compatible
"**taxonomic congruence"** requires that studies possess same set of taxa

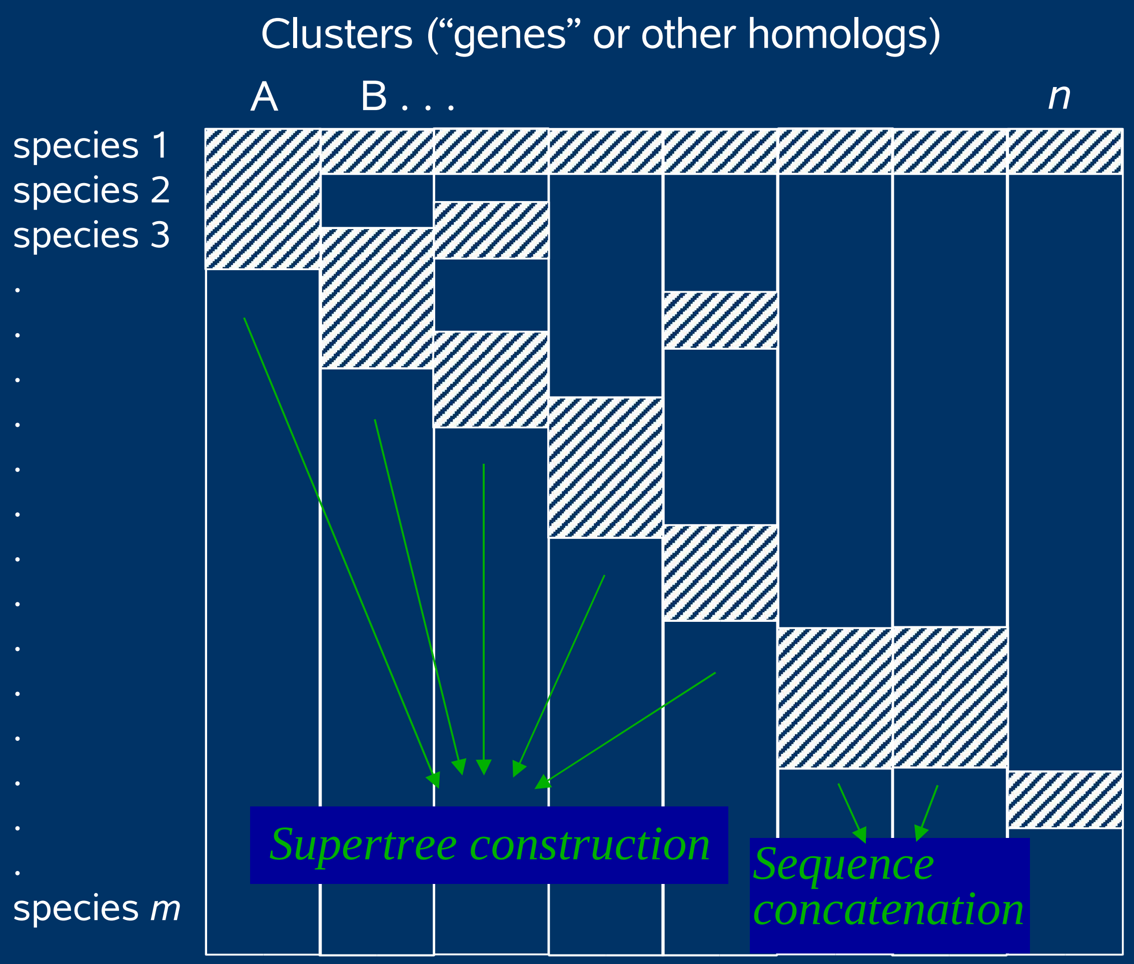
Some existing options

- **supermatrix** approach – combine original data sets into single, larger matrix
  - advantage: information retained in individual characters is useful
  - disadvantages:
    - gathering data to fill in gaps between taxa requires significant expense
    - some kinds of data cannot be included

- **concatenation of multiple sequences** from maximal number of taxa from sequence databases

- **supertrees** approach – estimates of phylogeny assembled from sets of smaller estimates (source trees) sharing some taxa but not necessarily all by combining trees *rather* than the data (Bininda-Emonds, 2004)

The sparse matrix of sequence and phylogenetic databases
(i.e., what we have NOW in databases)



Genbank release 127.0 (June 2003)

108,813 proteins from 11,5587 taxa (plants)

taxa x sequence clusters:
62 genes by 6 species
or
3 genes by 65 species

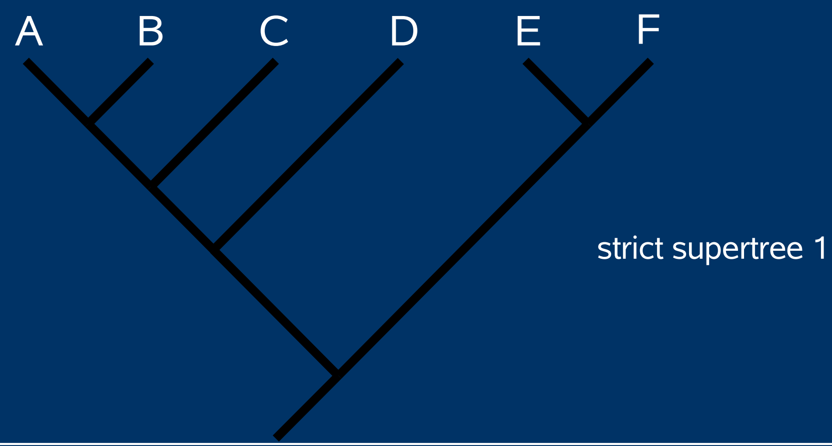
Data from Sanderson et al. (2003)

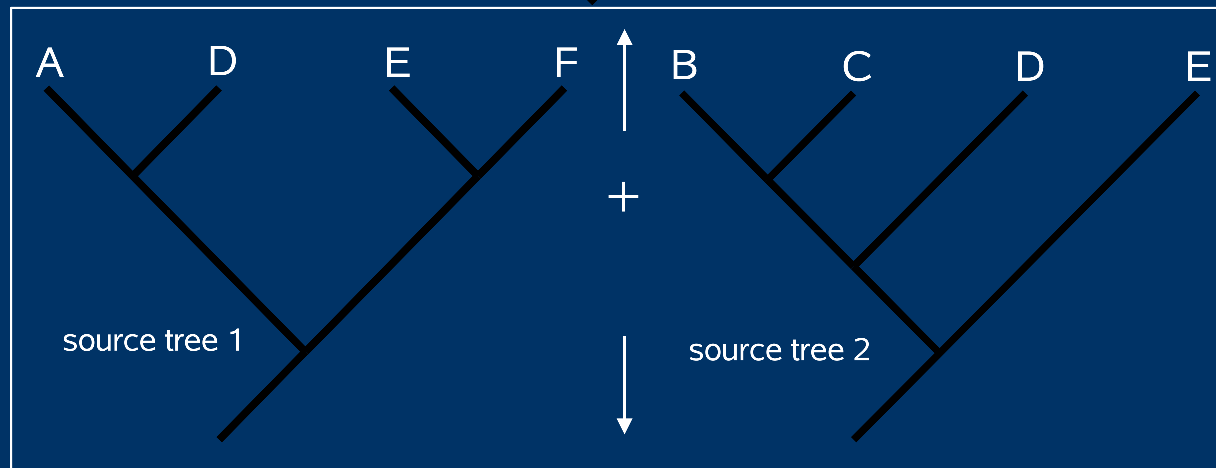
Supertree

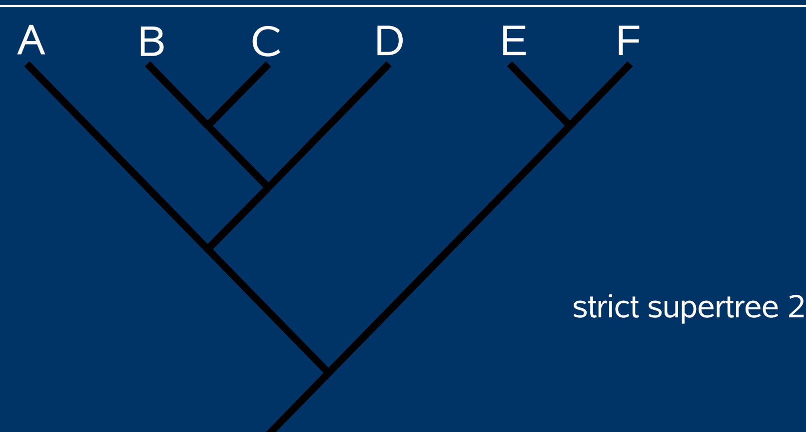
- Combination of phylogenetic trees that overlap taxonomically into a single larger tree using parsimony
  - Uses topologies of smaller trees rather than the actual data used to create those trees

Supertree terminology

Taxa found on only one source tree are **unique**; taxa found on two or more are **shared**. Any tree containing all the taxa found among the source trees is a **supertree**.





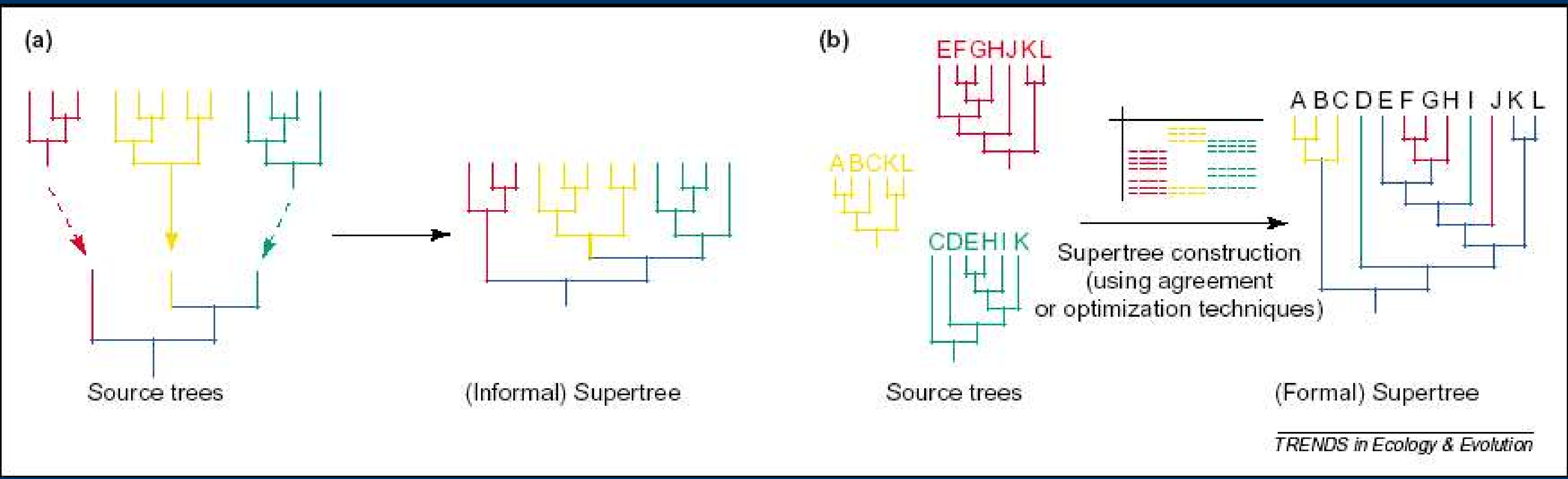


Two compatible **source** trees, together with two **strict** supertrees that are consistent with them despite disagreeing with each other.

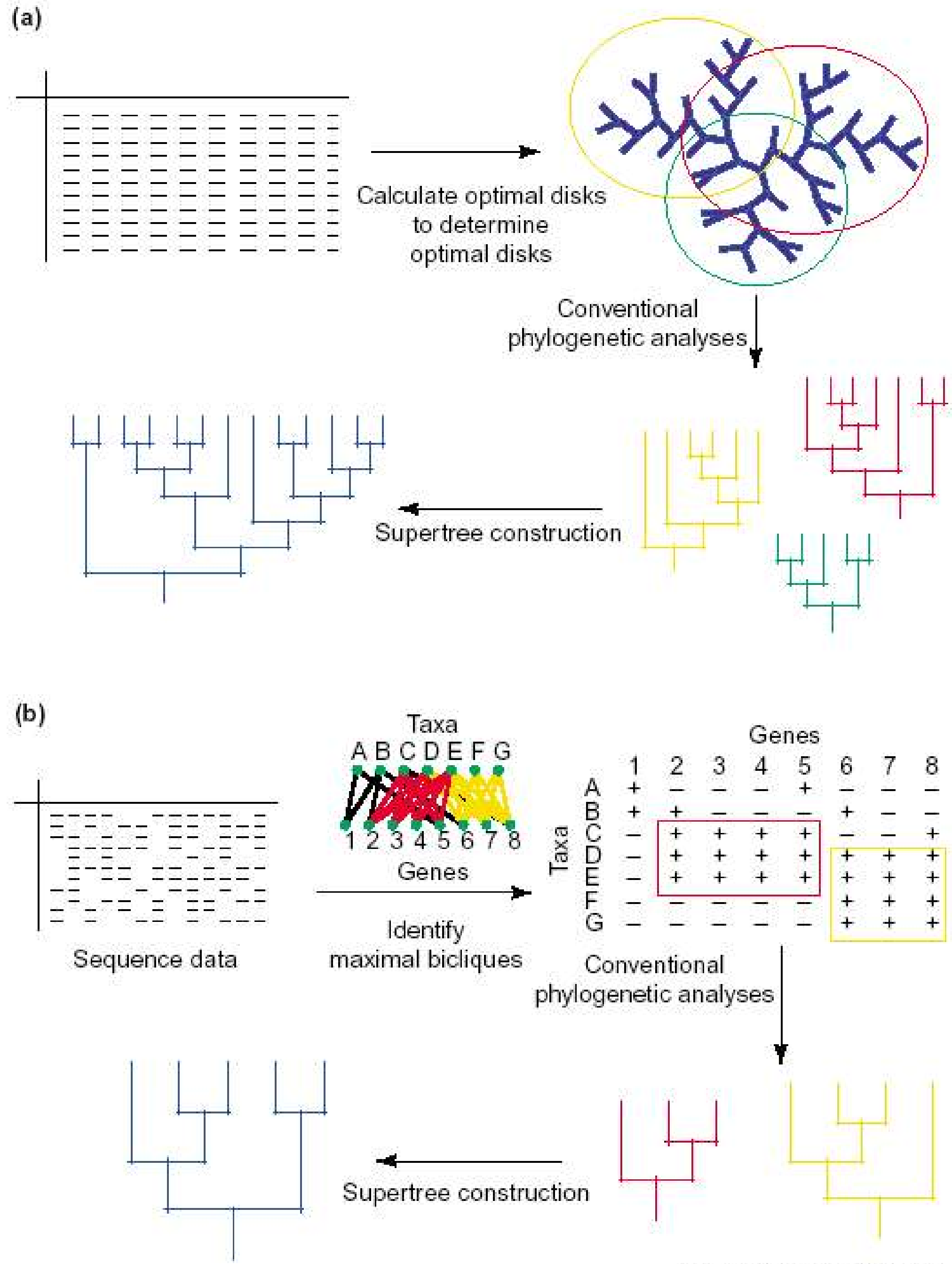
*From Sanderson et al. (1998)

Advantages of a Supertree

- allows phylogenetic estimates from all possible sources to be combined
- allows phylogenetic estimates from different kinds of analyses to be used
- combines estimates with different sets of terminal taxa to obtain a solution
- contains novel statements of relationship that are not present in any single source tree



(a)
Source trees
(Informal) Supertree
(b)
EFGHJKL
ABCKL
CDEHI K
Source trees
Supertree construction
(using agreement
or optimization techniques)
A B C D E F G H I J K L
(Formal) Supertree
TRENDS in Ecology & Evolution



TRENDS in Ecology & Evolution

Algorithms for Supertree Construction

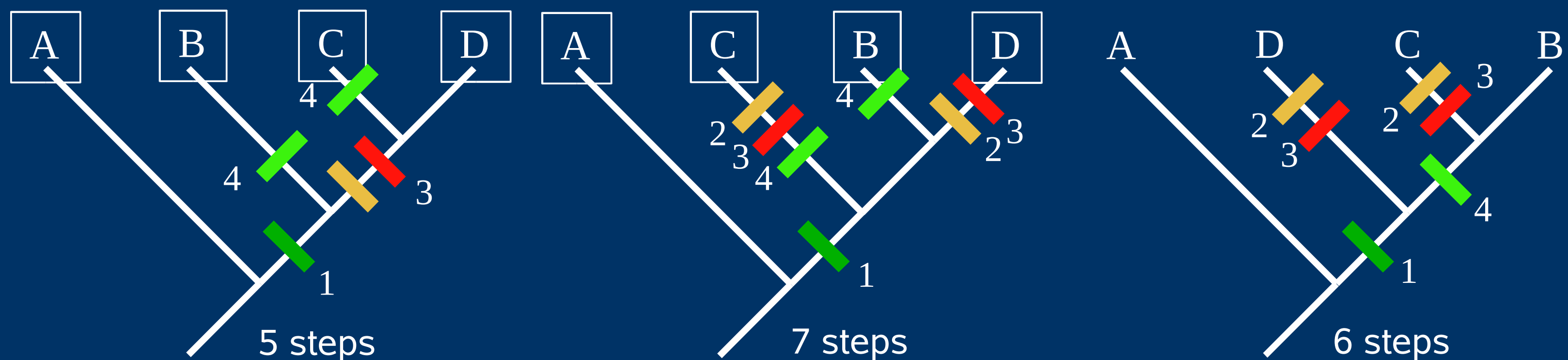
- Matrix Representation with Parsimony (MRP)
  - used whether or not source trees are compatible, or when there is conflict among source trees (esp. w/ large numbers)
  - method converts topology of each source tree into an equivalent data matrix representation, analysis using parsimony

- Strict Algorithm
  - used if source trees are compatible
  - tree construction is conservative and generally much faster than MRP

Parsimony

This data matrix contains character conflict. For example, character 4 suggests {B,C} is a monophyletic group, but characters 2 and 3 suggest {C,D} is monophyletic. They cannot both be true. How do we reconstruct phylogeny when the characters do not all agree?

	Characters			
Taxa	1	2	3	4
A	0	0	0	0
B	1	0	0	1
C	1	1	1	1
D	1	1	1	0

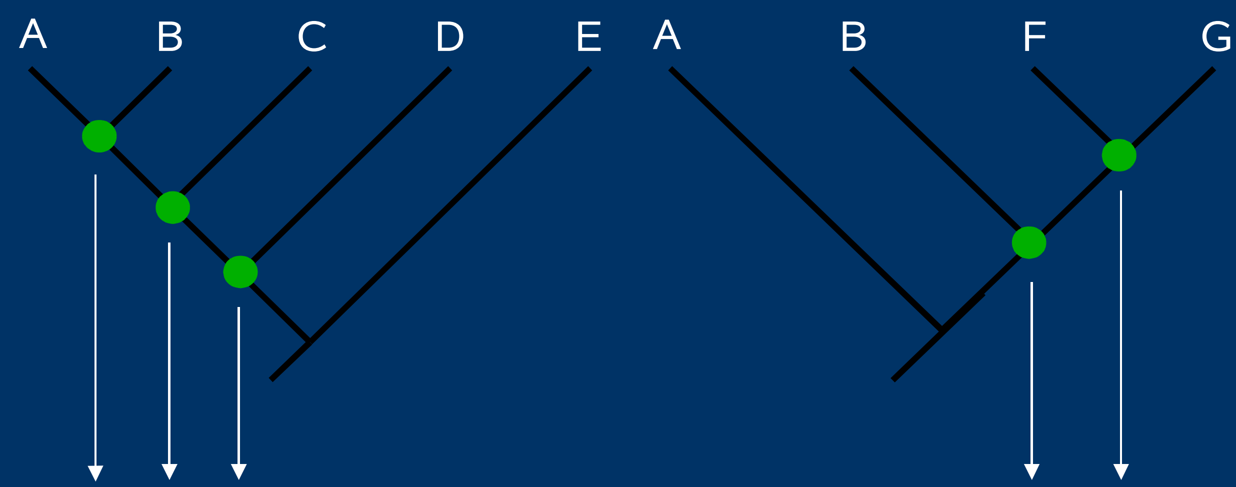


Phylogenetic analysis using parsimony is a procedure by which individual hypotheses of synapomorphy (shared, derived characters) are “tested” against one another for their overall explanatory power. The tree reconstruction with the fewest number of character state changes (sum of # of changes or **length**=5) is considered the most parsimonious of the three possible solutions.

Matrix Representation with Parsimony

In MRP a new matrix is constructed whose characters refer to the **topologies** of the source trees. Each clade (node) on a source tree yields one character in the matrix. Two schemes have been proposed for determining which taxa are scored as ‘0’, ‘1’, or ‘?’. Baum and Ragan scheme shown below:

Score ‘1’ for each taxon in clade, a ‘0’ for each taxon not in a clade, and a ‘?’ for taxa not present in that source tree. The characters from all source trees are then combined into one matrix and analyzed with parsimony. Trees then rooted with hypothetical ancestor having states with all ‘0’s.



A	1	1	1	. . . .	0	0
B	1	1	1	. . . .	1	0
C	0	1	1	. . . .	?	?
D	0	0	1	. . . .	?	?
E	0	0	0	. . . .	?	?
F	?	?	?	. . . .	1	1
G	?	?	?	. . . .	1	1

Table 1. Current formal supertree methods divided according to category

Agreement supertrees	Refs	Optimization supertrees	Refs
MinCutSupertree	[50]	Average consensus (matrix representation using distances, MRD)	[51]
Modified mincut supertree	[52]	Bayesian supertrees	[46]
RankedTree	[53]	Gene tree parsimony	[36]
Semi-Labelled- and AncestralBuild	[15]	Matrix representation using compatibility (MRC)	[38,54]
Semi-strict	[25,55]	Matrix representation using flipping (MRF; also known as MinFlip supertrees)	[26]
Strict	[7]	Matrix representation using parsimony (MRP) and variants	[10,11,24,54,56]
Strict consensus merger	[47]	Most similar supertree method (dfit)	[a]
		Quartet supertrees	[28,57]

Table 2. Examples of supertrees constructed using formal methods

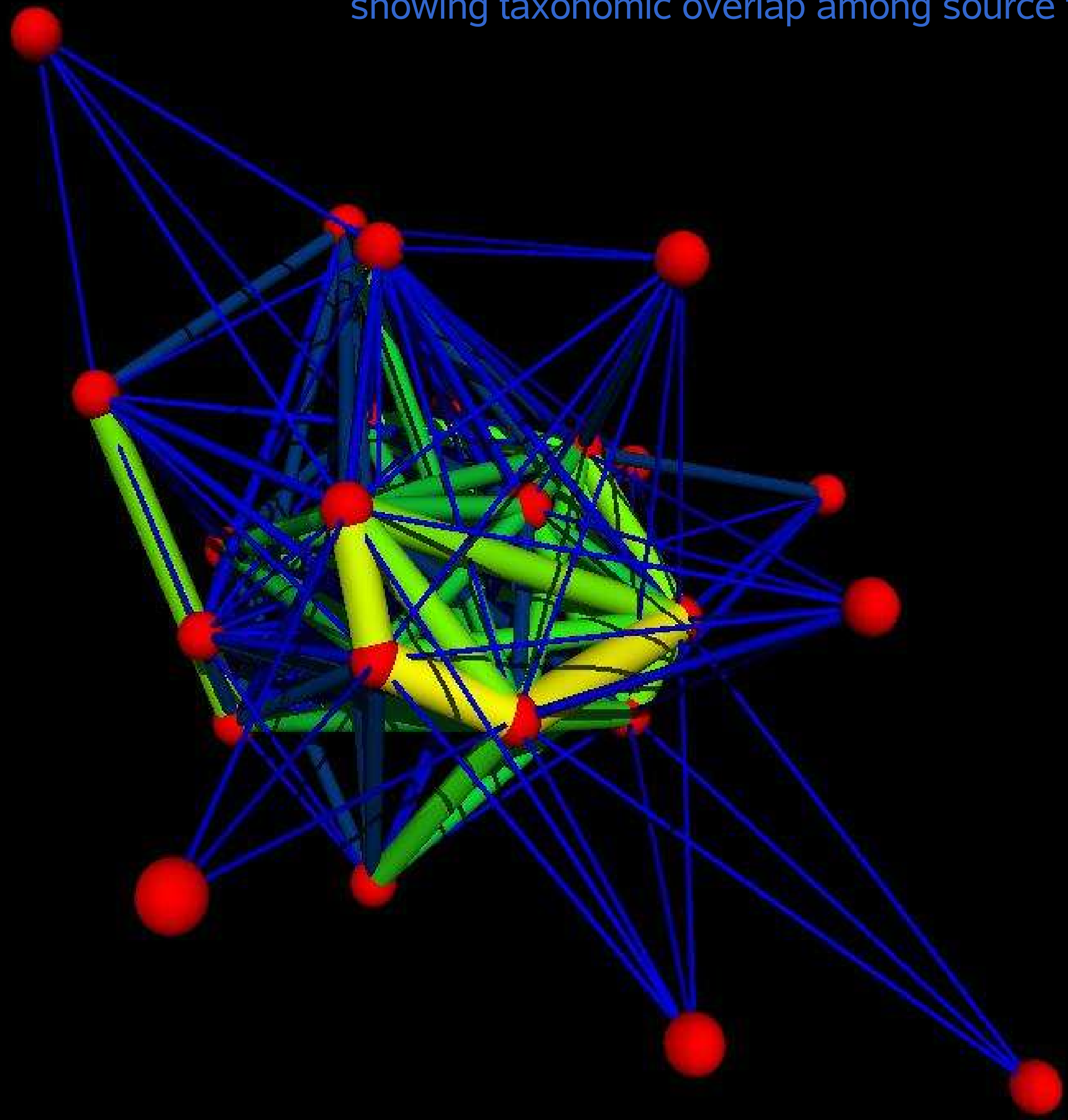
Group	Taxonomic level	No. terminal taxa[a]	Method[b]	No. source trees	Refs
Non-mammalian vertebrates					
Caenophidia (snakes)	Species	63	MRP	15	[58]
Crocodylia (crocodiles and relatives)	Species	**22 extant** + 53 fossil	MRP	21	[59]
Dinosauria (dinosaurs)	Genus	**277**	MRP	134	[60]
'Global avian fauna'	Genus and species	Not given	MRP/MRD/informal	90	[61]
Procellariiformes (seabirds)	Species	**122**	MRP	7	[34]
Mammals					
Artiodactyla (excl. whales) (even-toed ungulates)	Species	171	MRP	48	[62]
Carnivora (carnivores)	Species	**271**	MRP	177	[39]
Chiroptera (bats)	Species	**916**	MRP	105	[63]
Lipotyphla (insectivores)	Species	181	MRP	47	[64]
Lagomorpha (rabbits and pikas)	Species	**80**	MRP	146	[65]
Mammalia (mammals)	Order/Family	**90**	MRP	430	[30]
Marsupialia (marsupials)	Species	**267**	MRP	158	[66]
Primates (primates)	Species	**203**	MRP	112	[19,67]
Plants					
Angiosperms (flowering plants)	~Order	128	MRP	7	[68]
Angiosperms (flowering plants)	Family	379	MRP	46	[69]
Apiales (umbelliferous plants)	~Family	212	MRP	11	[68]
Cortaderia + outgroups (grasses)	Species	59	MRP	2	[70]
Hologalegina (legumes)	Species	571	MRP	43	[71]
Lithocarpus (tanbark oaks)	Species	22	MRP	5	[72]
Pinus (pines)	Species	**99**	MRP	14	[73]
Poaceae (grasses)	Genus	403	MRP	55	[74]
Other					
Bacteria	Phylum	9	MRD analogue	15	[75]
Bacteria	Species	37	MRP	130–196	[32]
Bacteria	Species	45	MRP	730	[33]
Diptera (true flies)	Family	**151**	MRP	12	[c]
Metazoa (animals)	'Class'	102	MRP	156	[76]
Schistosoma (blood flukes)	Species	14	MRP	8	[77]

[a]Entries in bold face are complete at the given taxonomic level for the clade in question.
[b]MRP, matrix representation using parsimony; MRD, matrix representation using distances; informal, informal supertree construction.
[c]David Yeates *et al.*; http://www.inhs.uiuc.edu/cee/therevid/supertree.html.

Literature Search

- Searched for published phylogenetic studies on Fabaceae Family (ISI Web of Science)
  - Keywords legumes, Fabaceae, systematics
  - Also searched for authors that have published in this field before
- Found 185 Studies published since 1984
- Studies used a variety of characters:
  - Gene sequences, non-coding DNA sequences, Morphology, binary characters (loss of chloroplast IR)

Example of a 'tree-graph' of phylogenies, showing taxonomic overlap among source trees.



(from Sanderson 2002)

Database

- Created an Access Database to store information on each study
  - Citation
  - Main Taxon
  - Number of Taxa
  - Outgroup
  - Character (sequence, morphological)
  - Phylogenetic Method (parsimony)
  - Support Value
  - Genbank/Treebase
  - Trees Presented
  - Independence
  - PDF file of paper

Trees

- Narrowed list
  - Eliminated studies with no taxonomic overlap (contained no taxa contained in another study)
  - Eliminated studies where primary data overlapped
  - Eliminated non-relevant studies
- Total # of candidate trees chosen = 68

Tree Descriptions

- Downloaded tree descriptions from Treebase (14)
- Wrote to authors and asked for tree descriptions (9) (Newick format)
- Had tree descriptions from a previous study (16)
- Made tree descriptions using MacClade (28)
- Unable to obtain (14)
- Opportunity to “edit”

Editing Tree Descriptions

- Naming Errors and Standardization
  - Misspellings, accession numbers
- Formatting Errors (trees from authors)
- Removing duplicate taxa or taxon names
  - Multiple accessions for the same species
- Synonomy
  - Multiple names for the same organism
  - Have not dealt with this issue yet

Tried Online Supertree Programs

- Rod Page's Supertree server (http://darwin.zoology.gla.ac.uk/cgi-bin/supertree.pl)
- Iowa State's Supertree server (http://genome.cs.iastate.edu/supertree/userdata_analysis/userdata_analysis.html)
- These sites have limitations

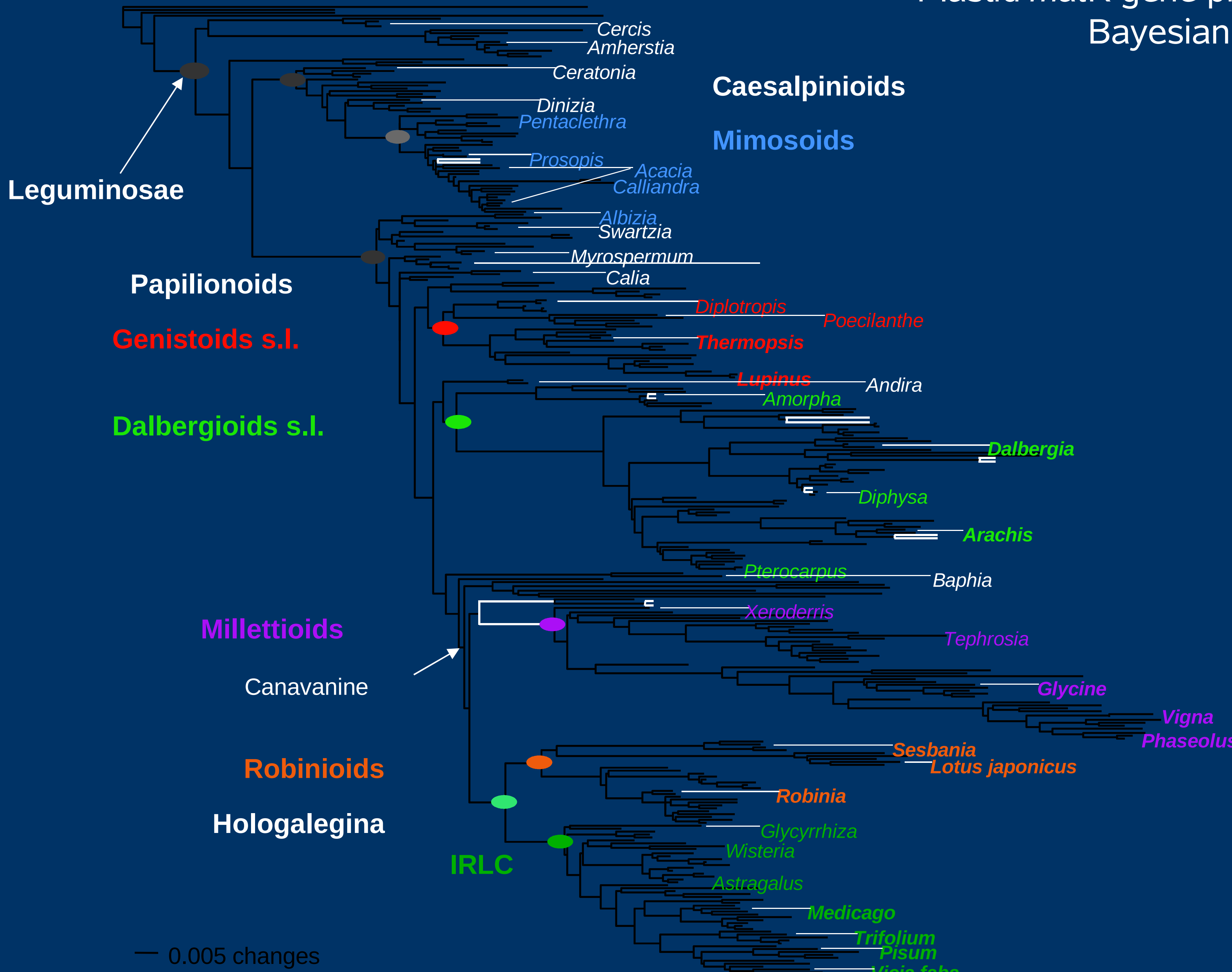
Creating Three Supertrees

- Break down project into manageable bits
- Divided the studies into subfamilies
  - Papilionoids
  - Mimosoids
  - Caesalpinioid
- Created a trees file for each group

Advantage

- Mimosoids and Papilionoids are monophyletic groups
- Typically the three groups are studied independently
- Each study has a different outgroup
  - Typically very distant and creates false relationships

Plastid *mat*K gene phylogeny
Bayesian analysis
330 taxa



Mimosoideae

- 3,000 species
- 58 genera

Albizia julibrissin Durazz.

Mimosoid Studies

•	2004	Wojciechowski M.F.	34/330 taxa
•	2003	Hughes C.E	72 taxa
•	2003	Miller J.T	60 taxa
•	2000	Clarke H.D	26 taxa

- Mimosoid Supermatrix 216 taxa, 429 characters

Caesalpinioideae

- 2,000 species
- 162 genera

Cercidium floridum Torr.

Caesalpinioid Studies

2004	Wojciechowski M.F.	33/330 taxa
2003	Haston E.M.	28 taxa
2003	Herendeen P.S.	220 taxa
2003	Schnabel A.	13 taxa
2003	Simpson B.B	81 taxa
2002	Davis C.C	7 taxa
2001	Brouat C.	13 taxa
1998	Schnabel A.	13 taxa

- Caesalpinioid Supermatrix 650 taxa, 602 characters

Papilionoideae

- Largest subfamily
  - 12,000+ species
  - 450 genera

Erythrina L.

Papilionoid Studies

- 2004 Wojciechowski M.F. 262/330 taxa
- 2004 Allan G.J 52 taxa
- 2004 McMahon M. 240 taxa
- 2004 Pardo C. 78 taxa
- 2004 Ree R. 15 taxa
- 2003 Ainoche A. 34 taxa
- 2003 Crisp M.D.66 taxa
- 2003 Dong T.X.X 10 taxa
- 2003 Kang Y. 56 taxa
- 2003 Lavin M. 12 taxa
- 2003 Schrire B.D. 109 taxa
- 2003 Steele K.P. 84 taxa
- 2002 Badr A. 37 taxa
- 2002 Cubas P. 57 taxa
- 2002 Doi K. 23 taxa
- 2002 Hu J-M 42 taxa
- 2002 Mayer 12 taxa
- 2002 Percy D.M. 50 taxa
- 2001 Bena G. 77 taxa
- 2001 Chandler G.T. 57 taxa
- 2001 Lavin M. 61 taxa
- 2001 Lavin M. 95 taxa
- 2001 Pennington R.T. 122 taxa
- 2000 Allan G.J. 42 taxa
- 2000 Crisp M.D.99 taxa
- 2000 Murphy D.J. 19 taxa
- 1999 Ainoche A-K 49 taxa
- 1999 Delgado-Salinas A. 132 taxa
- 1999 Wagstaff S.J. 39 taxa
- 1999 Wojciechowski M.F. 115 taxa
- 1998 Asmussen C.B. 42 taxa
- 1998 Bena G. 13 taxa
- 1998 Downie S.R. 62 taxa
- 1998 Fennel S.R. 10 taxa
- 1998 Lavin M. 34 taxa
- 1997 van Oss H. 8 taxa
- 1996 Sanderson M.J. 41 taxa
- 1995 Pennington R.T 27 taxa
- 1994 Liston A. 51 taxa
- 1993 Bruneau A. 66 taxa
- 1993 Doyle J.J. 53 taxa
- 1993 Sanderson M.J. 33 taxa
- 1992 Liston A. 64 taxa

Papilionoid Supermatrix 1502 taxa, 1683 characters

Create Supermatrix

- Used program R8S to create “supermatrix” from the trees file (Nexus output file)
- R8S is a program for estimating absolute rates of molecular evolution
- Used MRP algorithm
  - Matrix Representation with Parsimony

```
Vauquelinia_californica          000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Polygala_californica               0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Suriana_maritima                   1100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Quillaja_saponaria                 1100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Bauhinia_tomentosa              1011100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Cercis_gigantea                     101111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Cercis_occidentalis                1011110000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Cercis_canadensis                 10111100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Ceratonia_siliqua                    10100000000000111000000000000000000000000000000000000000000000000000000000000000000000000000000
Gymnocladus_chinensis          1010000000000011100000000000000000000000000000000010000000000000000000000000000000000000000000000000
Gleditsia_sinensis                   10100000000000111000000000000000000000000000000000110000000000000000000000000000000000000000000
Gleditsia_triacanthos              1010000000000011100000000000000000000000000000000011000000000000000000000000000000000000000000000
Arcoa_gonavensis                  10100000000000111000000000000000000000000000000000000000000000000000000000000000000000000000000000
Colophospermum_mopane     1011001100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Prioria_copaifera                     1011001100000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Hymenaea_courbaril               101100101100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Tessmannia_lescrauwaetii       1011001011000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Brownea_sp                            1011001010110000000000000000000000000000000000000000000000000000000000000000000000000000000000
Oddoniodendron_micranthum 1011001010111100000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Berlinia_congolensis               1011001010111100000000000000000000000000000000000000000000000000000000000000000000000000000000000
Brachystegia_spiciformis         101100101011100000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Cynometra_mannii                  10110010101100000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Amherstia_nobilis                    10110010101000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Petalostylis_labicheoides        101000000000001000000000000000000000000000000000000000000000000000000000000000000000000000000000000
Dialium_guianensis                 1010000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000
Erythrostemon_gilliesii            1010000000000011111000000000000000000000000000000000000000000000000000000000000000000000000000000
Caesalpinia_andamanica         1010000000000011111100000000000000000000000000000000000000000000000000000000000000000000000000000000
Caesalpinia_pulcherrima         ????????????????????????????????????????????????????????????????????????????????????????????????????
Haematoxylum_brasiletto         1010000000000011111100000000000000000000000000000000000000000000000000000000000000000000000000000000
Chamaecrista_fasciculata        1010000000000011110010000000000000000000000000000000000000000000000000000000000000000000000000000000
Senna_candolleana                 10100000000000111100110000000000000000000000000000000000000000000000000000000000000000000000000000
Senna_covesii                         1010000000000011110011000000000000000000000000000000000000000000000000000000000000000000000000
Peltophorum_dubium              1010000000000011110000111000000000000000000000000000000000000000000000000000000000000000000000000000
Cercidium_floridum                 1010000000000011110000111110000000000000000000000000000000000000000000000000000000000000000000000
Parkinsonia_aculeata              10100000000000111100001111100000000000000000000000000000000000000000000000000000000000000000000000
Conzattia_multiflora                101000000000001111000011110000000000000000000000000000000000000000000000000000000000000000000000
Poeppigia_procera                  1010000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000000
Dinizia_excelsa                       1010000000000011110000110000000000000000000000000000000000000000000000000000000000000000000000
Inga_punctata                          101000000000001111000010000111111111100000000000000000000000000000000000000000000000000000000
Samanea_saman                    101000000000001111000010000111111111000000000000000000000000000000000000000000000000000000000000000
Enterolobium_cyclocarpum     1010000000000011110000100001111111110000000000000000000000000000000000000000000000000000000000000000
Enterolobium_contortisiliquum ????????????????????????????????????????????????????????????????????????????????????????????????????
Lysiloma_watsonii                    10100000000000111100001000011111111001000000000000000000000000000000000000000000000000000000000
Lysiloma_acapulcensis            ???????????????????????????????????????????????????????????????????????????????????????????????????
Lysiloma_tergemina                 ?????????????????????????????????????????????????????????????????????????????????????????????????
Havardia_pallens                      ??????????????????????????????????????????????????????????????????????????????????????????????
Havardia_albicans                    ???????????????????????????????????????????????????????????????????????????????????????????????
```

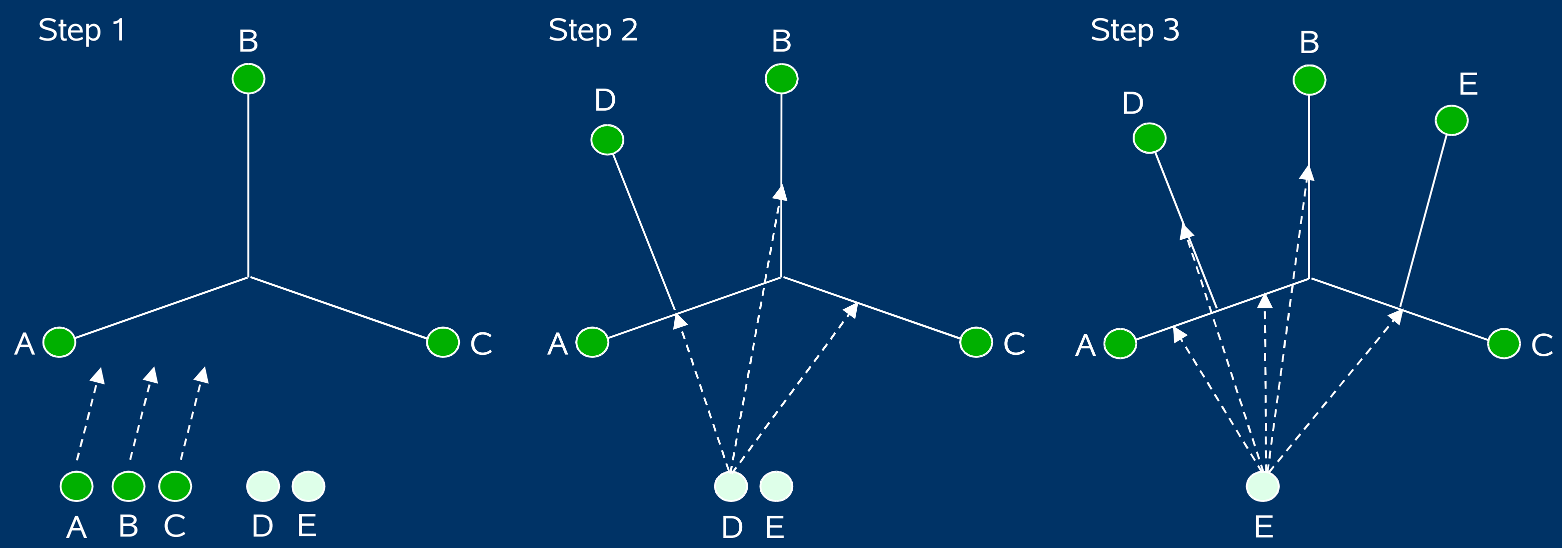
Topological Constraints

- Weighted characters in the supermatrix and member of the Fabaceae family and the Mimosoid subfamily as these are supported monophyletic groups

Heuristic Search

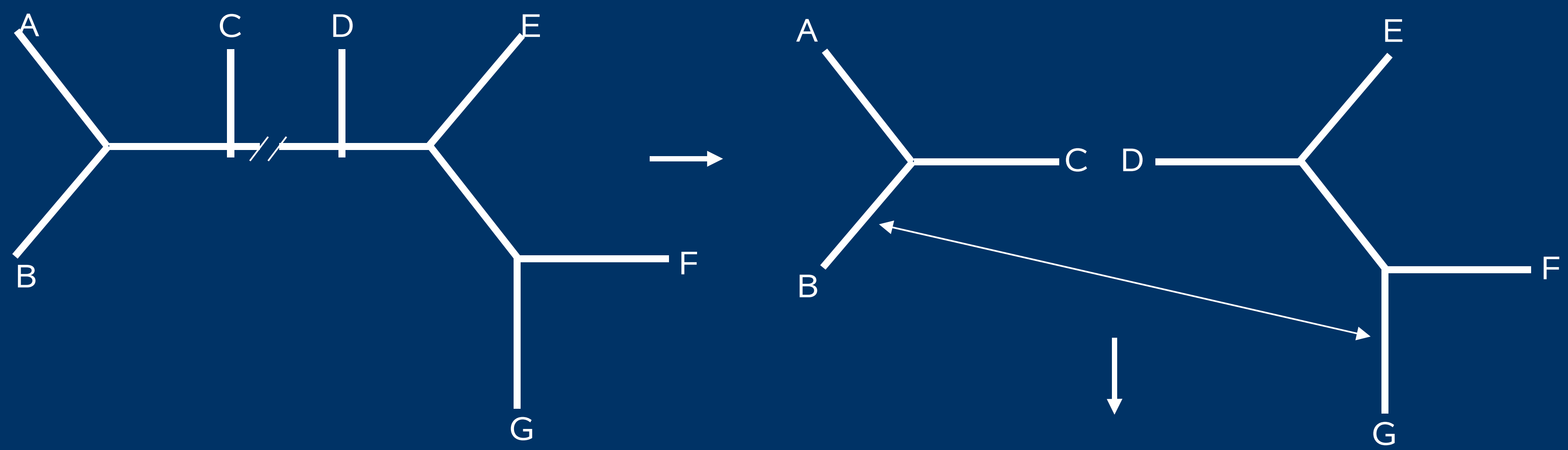
- Executed Supermatrix in PAUP software
  - Phylogenetic Analysis Using Parsimony
- Ran heuristic searches
  - storing 5000 trees maximum
  - holding five trees at each step
  - using the TBR (tree bisection-reconnection) branch-swapping algorithm
- 3 types of searches using different addition sequence procedures: simple, closest, random

Heuristic methods: step 1, making initial tree, taxon addition sequence

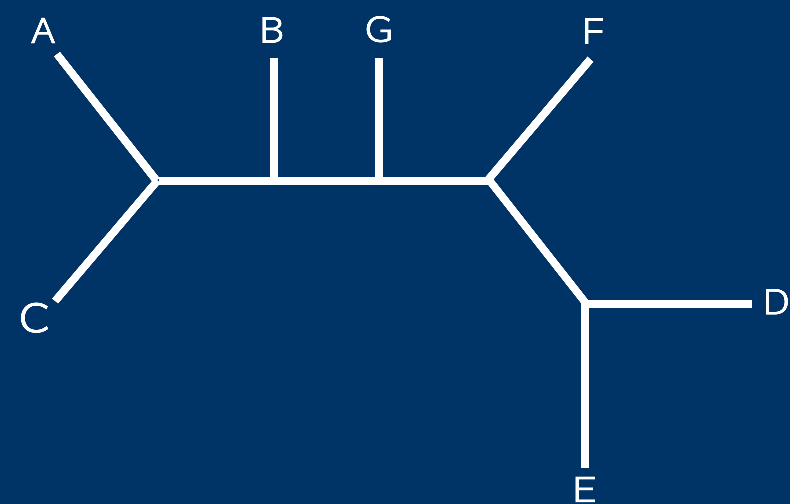


Taxa are always added sequentially to make a tree in this phase. The simplest order of addition is known as “ASIS” addition; here taxa are added in the order they appear in the matrix. The first three taxa are joined into an unrooted three-taxon tree, then the fourth taxon in the matrix is added. It can be added in one of three places, so the length of the tree is determined for each possibility and the placement that is optimal at that point in time is selected. Next, the fifth taxon is added, and so on, until a complete tree is built. Other addition sequence implemented in software such as *PAUP** include RANDOM (random order addition) and CLOSEST (which chooses next taxon to be added by finding the one that would add the fewest number of steps to the new tree).

Heuristic methods: step 2, branch swapping



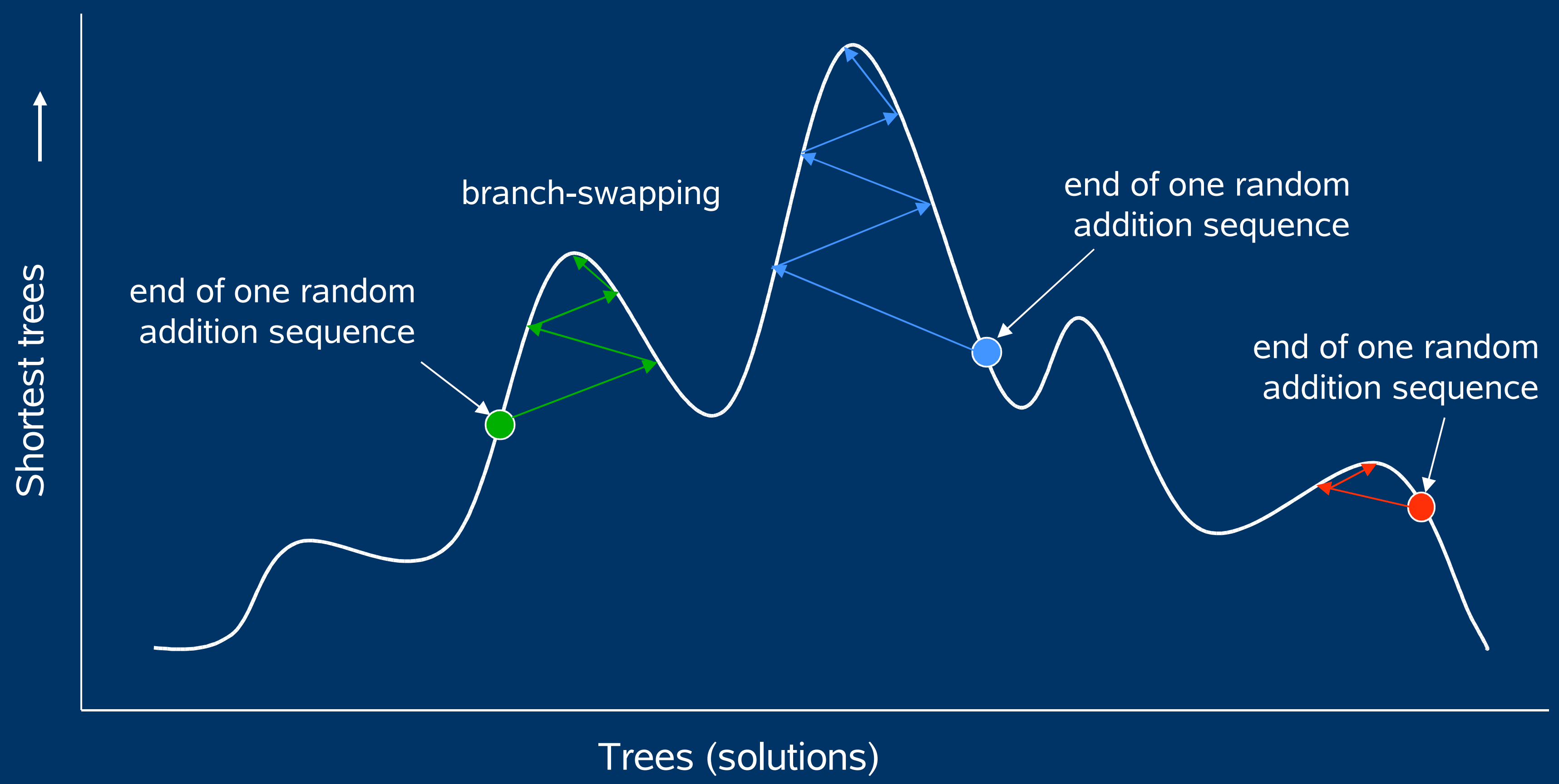
Branch swapping by tree bisection and reconnection (TBR). The tree is initially bisected along a branch, yielding two disjoint subtrees. The subtrees are then reconnected by joining a pair of branches, one from each subtree, with all possible bisections and reconnections evaluated. The shortest is saved and branch swapping proceeds again until a shorter tree is found.



(after Swofford et al. 1996)

Optimization methods

On a landscape of trees, random addition sequences (tree-building) are used to find multiple optima, or '**tree islands**'. Branch swapping moves search nearer to top of local optima. New random addition sequences may find additional local optima.



Consensus Tree

- Allowed search to find the maximum of 5000 trees for each heuristic search
- Created a 90% majority rule consensus tree for each of the heuristic searches
  - Rooted the tree with an outgroup
  - included all other compatible groupings

Mimosoid Supertree

90% Majority Rule TBR/5

Future Work

- Finish the supertrees for the Papilionoids
- Obtain remaining studies from authors and add to supertrees
- Combine the three supertrees into one super-supertree
- Compare this to work at UC Davis

References

- Bininda-Emonds, O. R. P. 2004. The evolution of supertrees. *Trends in Ecology and Evolution* 19:315-322.
- Bininda-Emonds, O. R.P. et al. 2004. Garbage in, garbage out: data issues in supertree construction. Chapter 12 in *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life. Computational Biology* 3:267-280.
- Pennisi, E. 2003. Modernizing the Tree of Life. *Science* 300: 1692-1697.
- Sanderson, M. J., A. C. Driskell, R. H. Ree, O. Eulenstein, and S. Langley. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Molecular Biology and Evolution* 20: 1036-1042.
- Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M. (1996). Phylogenetic Inference. In *Molecular systematics*, 2nd edition, chap. 5, pp. 407-514. Sinauer and Associates, Sunderland, Massachusetts

Acknowledgements

- Dr. Marty Wojciechowski
- Dr. Rosemary Renaut
- Dr. Bradford Kirkman-Liff