

Trees to Supertrees

When?

What?

How?

Issues?

When? What?

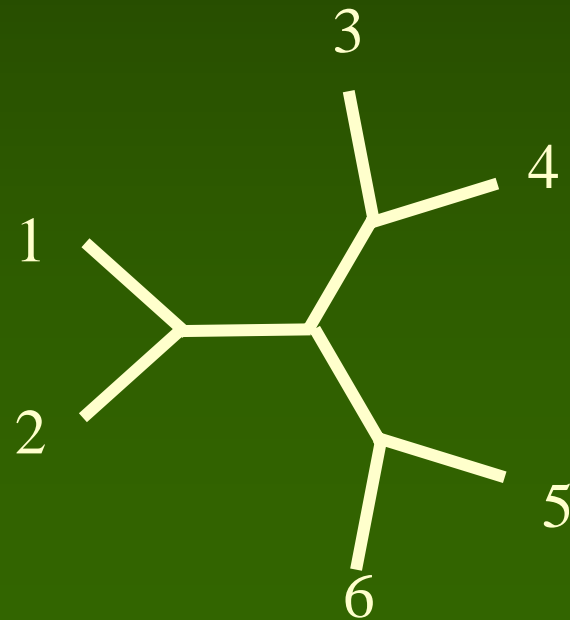
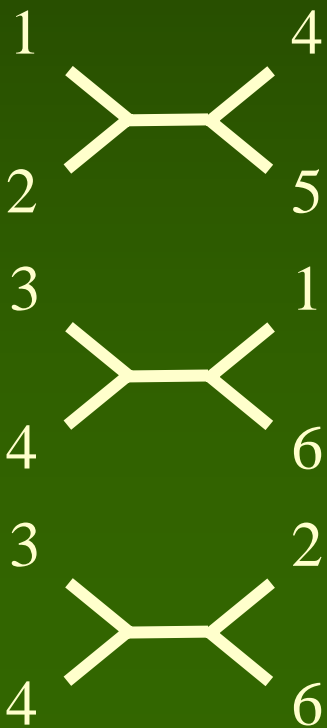
- As part of a phylogenetic analysis
 - Combine data from different sources
 - genome partitions (gene trees)
 - different character sets
 - different methodologies
- Tree of life
- Cospeciation studies
- Biogeographic studies

Elements and concepts

- Trees
 - Compatible
 - Incompatible
- Taxa
 - Identical sets
 - Inclusive sets
 - Overlapping sets
 - (exclusive sets)
- Data
 - Same
 - Partitioned
 - Different

Some terminology

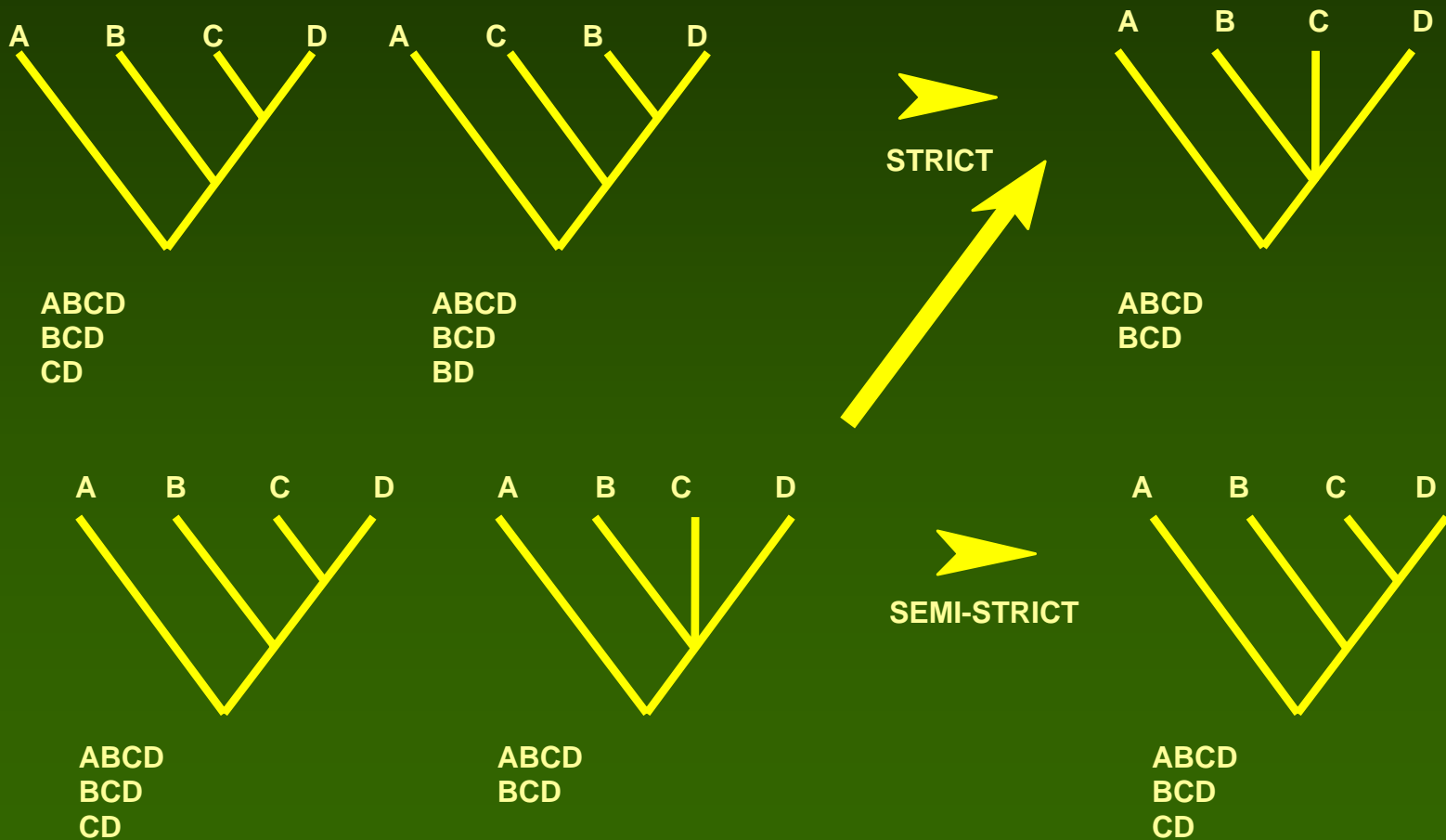
- A set of trees is called *compatible* if a (fully resolved) tree exists from which all other trees can be derived by removing some taxa (and their branches): This is the “parent tree”



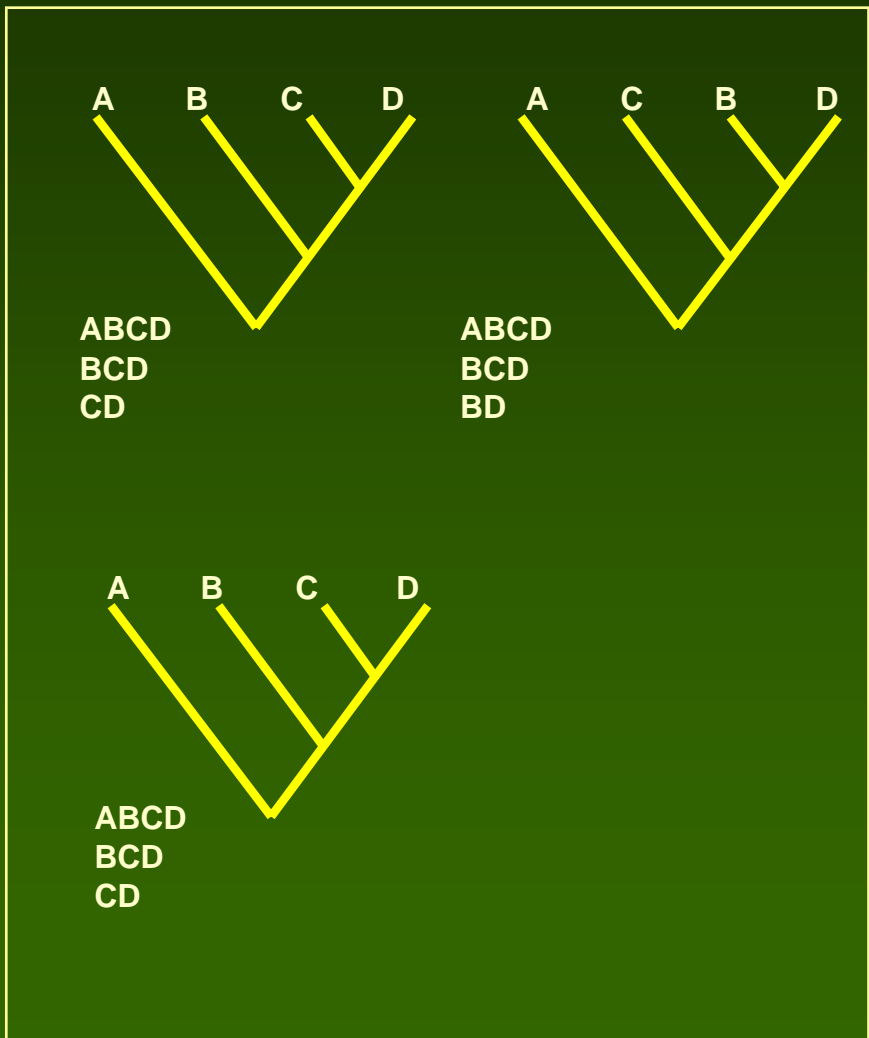
Methods

- Consensus methods: applicable when datasets are identical
 - Strict
 - Semi-strict
 - Majority
 - Adams
- Supertree methods: applicable also when they don't (but...)
 - Matrix-based methods
 - MRP
 - Semi strict
 - Flip
 - Non-matrix-based methods
 - MinCut (incl.modified Min Cut)
 - (Step matrix method)

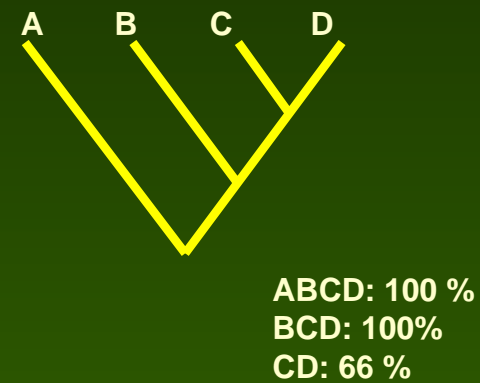
Strict and semi-strict consensus



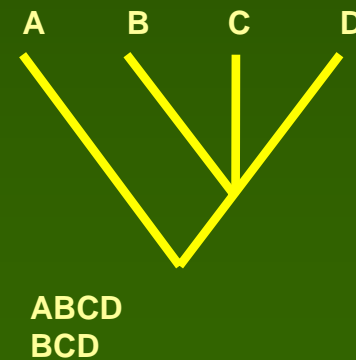
Majority Rule



Majority rule

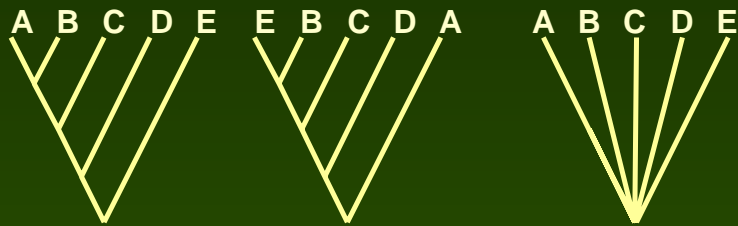


strict



Strict vs Adams consensus

Strict consensus



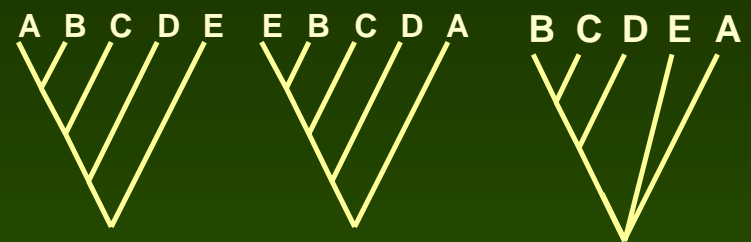
ABCDE
 ABCD
 ABC
 AB

ABCDE

EBCD
 EBC
 EB

ABCDE

Adams consensus



((ABCD)E)

((EBCD)A)

((ABC)DE)

((EBC)DA)

((ABC)D)

((EBC)D)

((AB)CDE)

((EB)CDA)

((AB)CD)

((EB)CD)

((AB)C)

((EB)C)

((BC)ADE)

((BC)ADE)

((BC)ADE))

((BC)AE)

((BC)AE)

((BC)AE)

((BC)AD)

((BC)AD)

((BC)AD)

((BC)E)

((BC)A)

((BC)D)

((BC)D)

((BC)D)

...

...

The Adams rule:

set A nests inside set B *iff* (if and only if)

(1) A is a subset of B, *and*

(2) the leaves in set A have a more recent common ancestor than the leaves in set B.

(from Page, 1993)

Supertree methods

Limitations

- Compatible trees: no problem
 - Any method for Supertree construction should aim to recover the Parent tree or all such trees
- Incompatible trees: various problems!
 - Lack of resolution
 - Creation of new clades

Matrix methods: common elements

- Recode each “fundamental” cladogram as a complex additive character
 - Coding strategies:
 - Standard: use question marks if taxa are unknown (absent) for one or more of the source trees
 - Modified “Purvis” coding: use question marks also to express implied relations
- Combine all these “characters” in a single matrix
- Use this matrix to compute a new cladogram

Additive binary coding

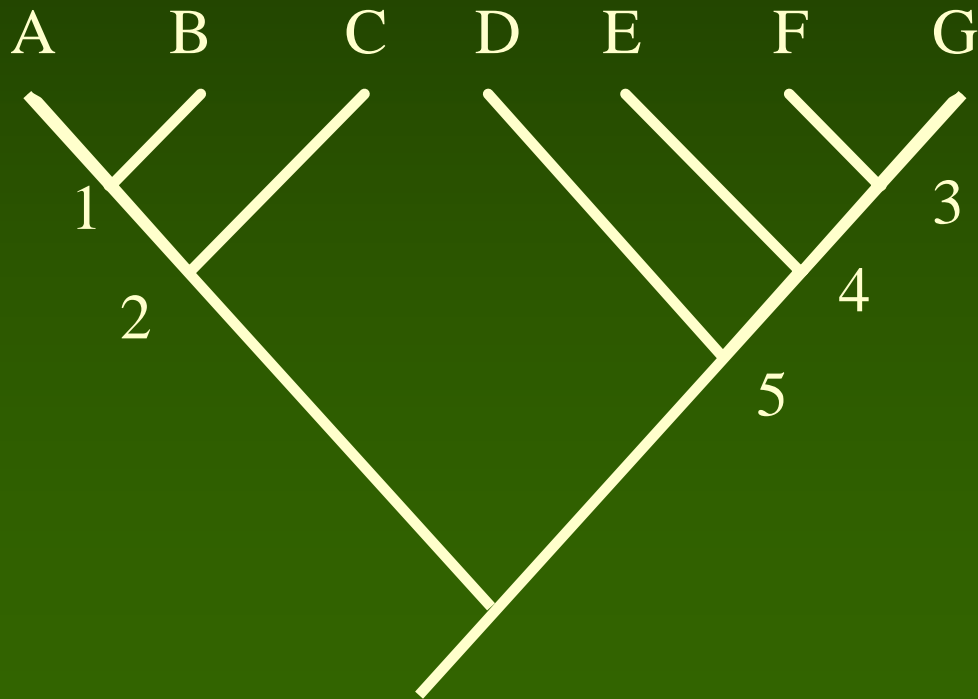
For characters:

- Arrange character states in a character state tree (“transformation series”)
- Code data columns in such a way that
- When the original character states are analyzed, the original character state tree is the result

For cladograms

- Do this for an entire cladogram

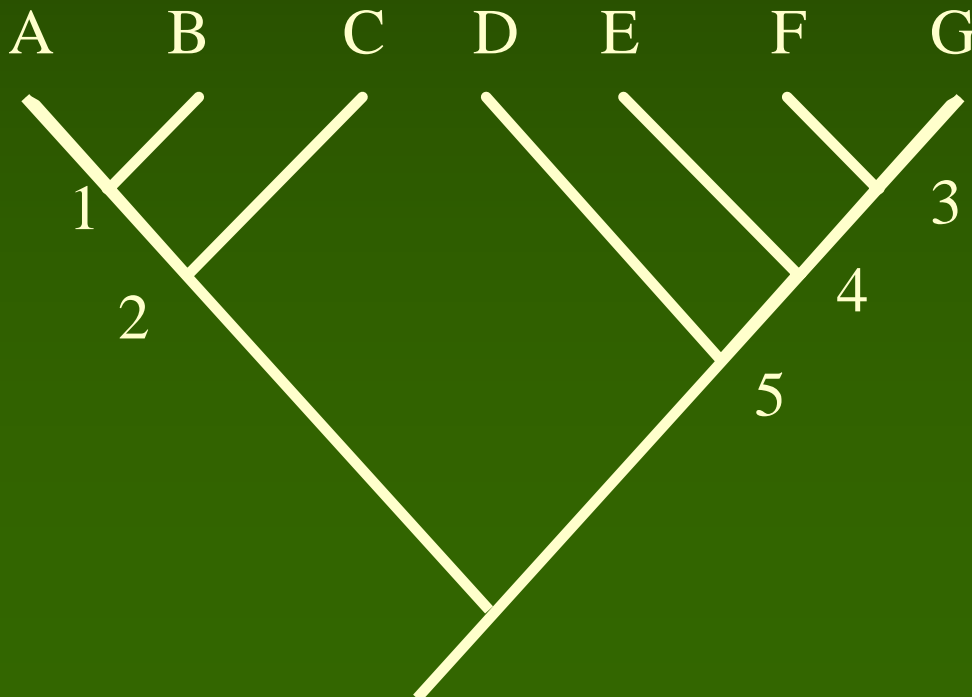
Additive binary coding



	1	2	3	4	5
A	1	1	0	0	0
B	1	1	0	0	0
C	0	1	0	0	0
D	0	0	0	0	1
E	0	0	0	1	1
F	0	0	1	1	1
G	0	0	1	1	1

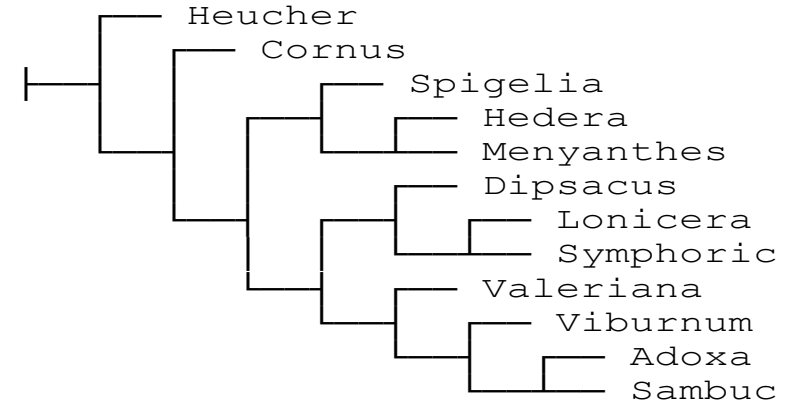
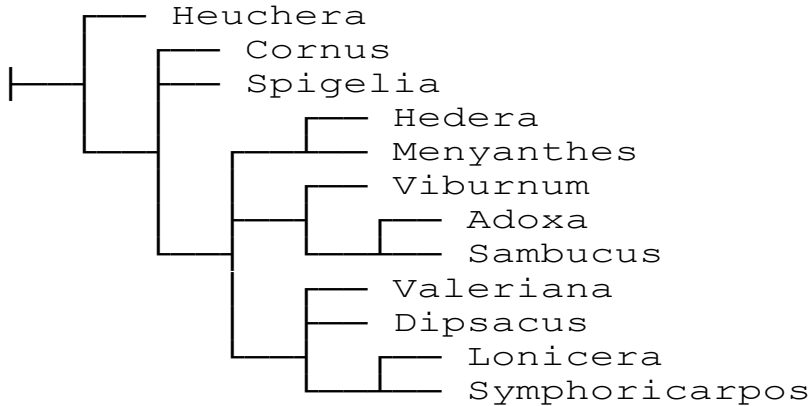
Additive binary coding: modification

- Purvis (1995): Additive binary coding introduces “redundancy”:
 - If (FG) is not a member of (ABC), the information that it is not a member of (AB) is redundant
 - Remove “redundancy” by introducing question marks



	1	2	3	4	5
A	1	1	?	?	0
B	1	1	?	?	0
C	0	1	?	?	0
D	?	0	?	0	1
E	?	0	0	1	1
F	?	0	1	1	1
G	?	0	1	1	1

Combine data into single matrix



- Once different trees are coded, the data can easily be combined into a single datamatrix

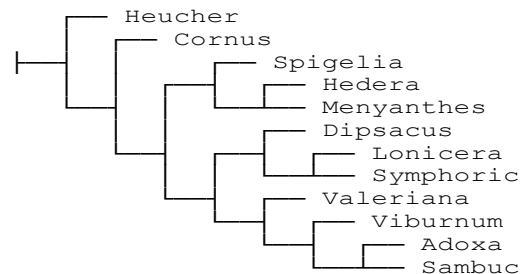
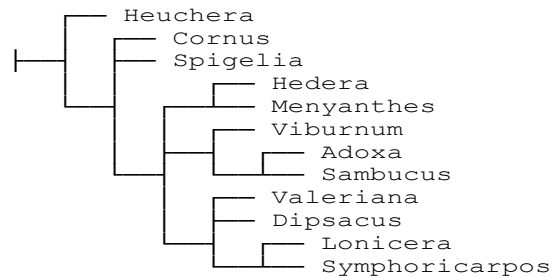
- All taxa have one row
- Enter “?” when absent from any single tree

ROOT	00000000	0000000000
Heucher	00000001	0000000000
Cornus	00100001	0001000000
Spigelia	00100001	0111000000
Hedera	11100001	1111000000
Menyanthes	11100001	1111000000
Valeriana	01100011	0011001100
Viburnum	01101001	0011011100
Adoxa	01111001	0011111100
Sambucus	01111001	0011111100
Dipsacus	01100011	0011000101
Lonicera	01100111	0011000111
Symphoricarp	01100111	0011000111

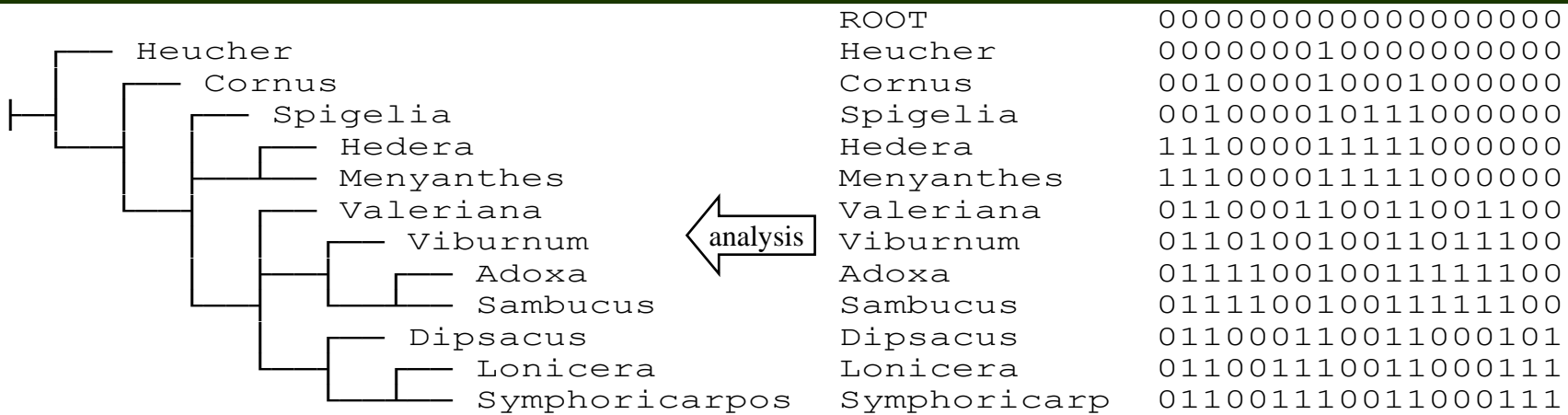
Conflict in combined matrix

- Combining incompatible trees will introduce conflict in the combined matrix
- How to deal with such character conflict:
 - Analyze with standard parsimony: minimize length
 - Analyze by minimizing “flips”: minimize “noise”
 - Reject contradictory characters: semi-strict supertrees

ROOT	00000000000000000000
Heucher	00000001000000000000
Cornus	00100001000100000000
Spigelia	00100001011100000000
Hedera	11100001111100000000
Menyanthes	11100001111100000000
Valeriana	C11000110011001100
Viburnum	C11010010011011100
Adoxa	C11110010011111100
Sambucus	C11110010011111100
Dipsacus	C11000110011000101
Lonicera	C11001110011000111
Symphoricarp	C11001110011000111



Parsimony



- Resolve contradiction using parsimony:
 - Minimizes change by invoking a common origin for as many changes as possible
 - Useful in an evolutionary context
 - But is this an evolutionary problem?

Flips?

Flip Supertree construction: Eulenstein et al. 2004; Syst. Biol. 53

- Do not minimize *total length*, but *flips*
 - single changes in the datamatrix (1 \leftrightarrow 0)
- Tree is selected that requires minimum number of flips
 - *Not dependent on evolutionary context*, therefore more suited for supertree analysis than standard parsimony?
 - Minimizes “noise”

Flips: minimizing “noise”

- One change from 1 to 0 resolves conflict in this character

ROOT	00000000000000000000
Heucher	00000001000000000000
Cornus	00100001000100000000
Spigelia	00100001011100000000
Hedera	11100001111110000000
Menyanthes	11100001111110000000
Valeriana	C11000110011001100
Viburnum	C11010010011011100
Adoxa	C11110010011111100
Sambucus	C11110010011111100
Dipsacus	C11000110011000101
Lonicera	C11001110011000111
Symphoricarp	C11001110011000111

Reject characters?

Semi-strict supertrees: Goloboff & Pol, 2002; *Cladistics* 18 (5)

- Use only characters that are not contradicted
 - Largest set: superclique
- Analyze with parsimony
- Results in strict consensus when trees have the same taxa
- But can be used for different taxon compositions

ROOT	00000000000000000000
Heucher	000000010000000000
Cornus	001000010001000000
Spigelia	001000010111000000
Hedera	111000011111000000
Menyanthes	111000011111000000
Valeriana	C11000110011001100
Viburnum	C11010010011011100
Adoxa	C11110010011111100
Sambucus	C11110010011111100
Dipsacus	C11000110011000101
Lonicera	C11001110011000111
Symphoricarp	C11001110011000111

Matrix methods:

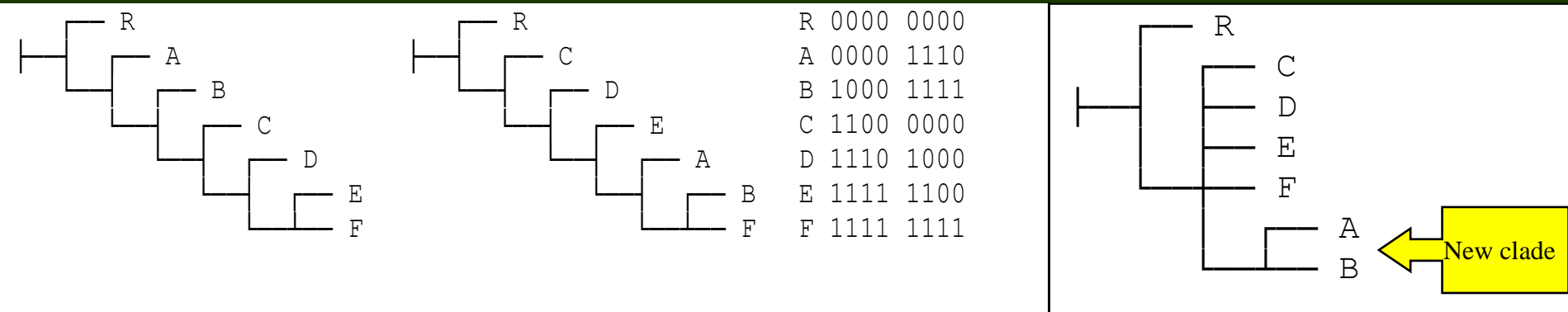
– Advantages

- Can often easily be implemented using standard software

– Disadvantages

- Search time increases with number of taxa, comparable to standard phylogenetic analysis
- Therefore: no significant increase in computing speed for large numbers of taxa compared to a combined analysis
- Creation of “new” clades

MRP: new clades

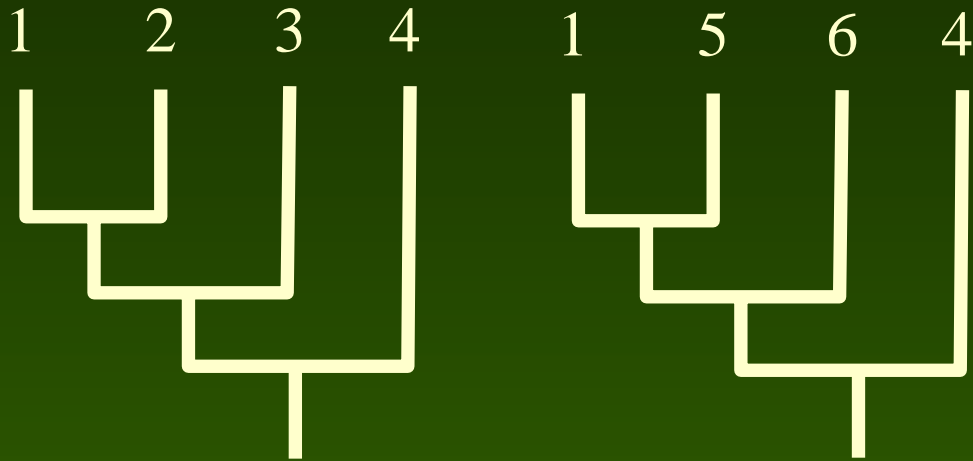


- Goloboff & Pol, 2002. *Cladistics* 18 (5).
- MRP can produce new clades
 - not present in any source tree
 - contradicted in all source trees
- Based on reversal of “character states” interpreted as synapomorphy

Non matrix-based methods

- MinCut algorithm
- Step-coding

MinCut algorithm



Two compatible trees

①

②

⑥

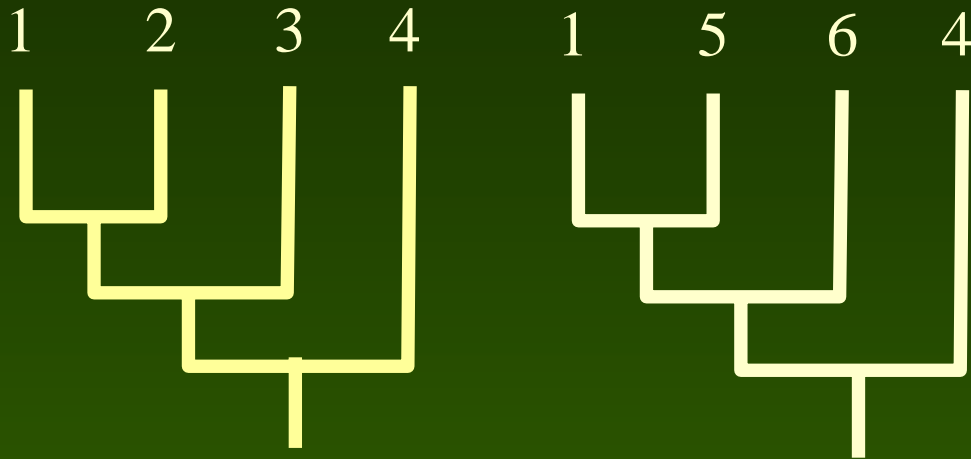
③

⑤

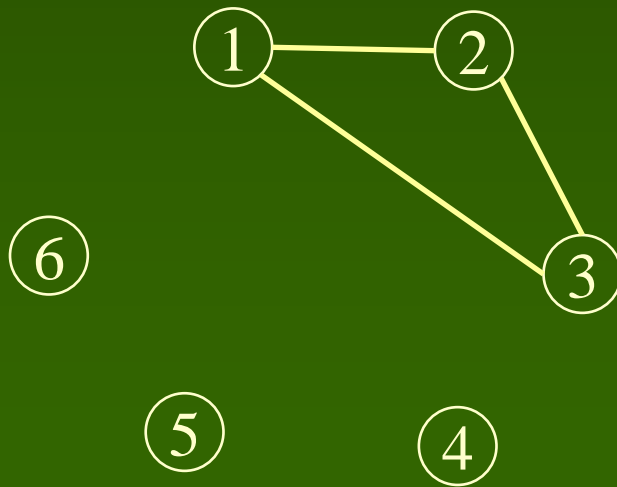
④

Display as graphs:

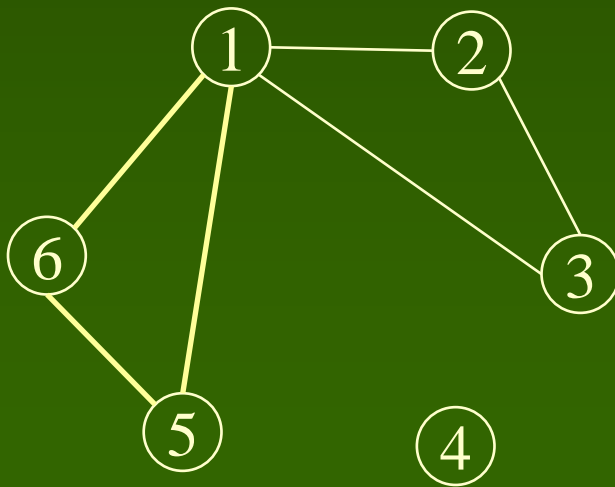
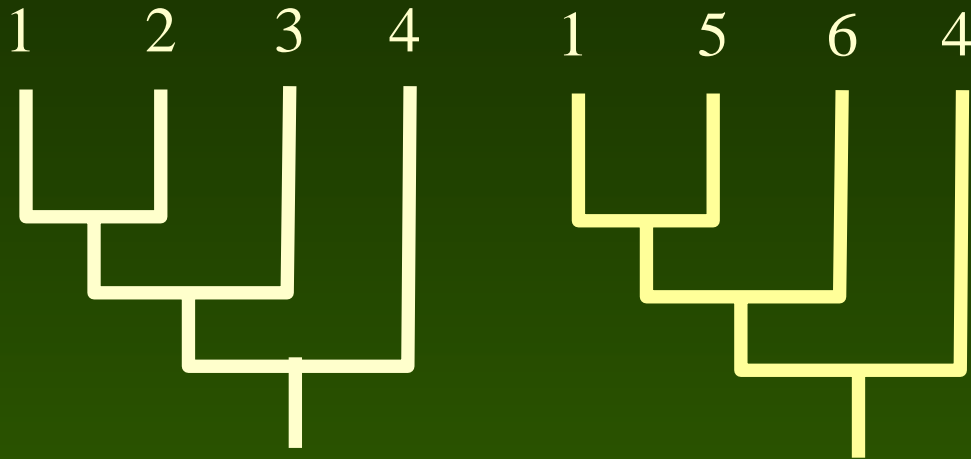
MinCut algorithm



- Using the graphs:
- Form clusters:
 - Connect two taxa if they are (nontrivially) clustered in at least one of the trees

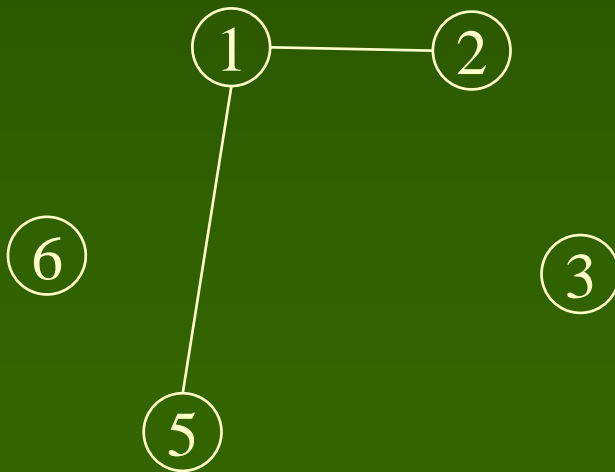
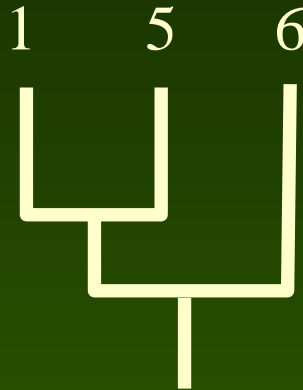
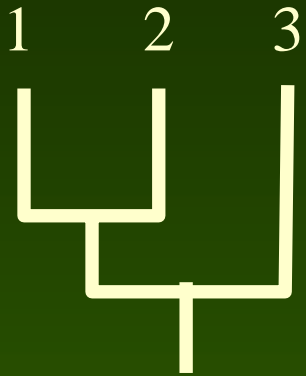


MinCut algorithm



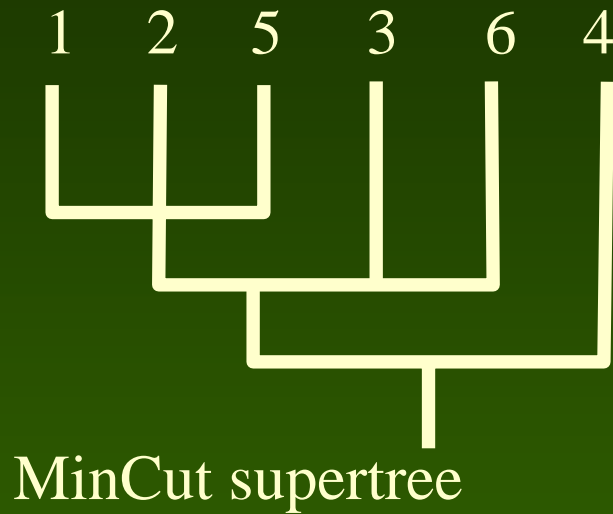
- Using the graphs:
- Form clusters:
 - Connect two taxa if they are (nontrivially) clustered in at least one of the trees
- Retain cluster: (1 2 3 5 6)
- Discard isolated taxa

MinCut algorithm



- (collapse taxa if joined in all cladograms)
- Reduce trees to contain only the taxa from the connected cluster
- Continue as before
- Cluster found: (1 2 5)
- Repeat until no more clusters are found

MinCut algorithm

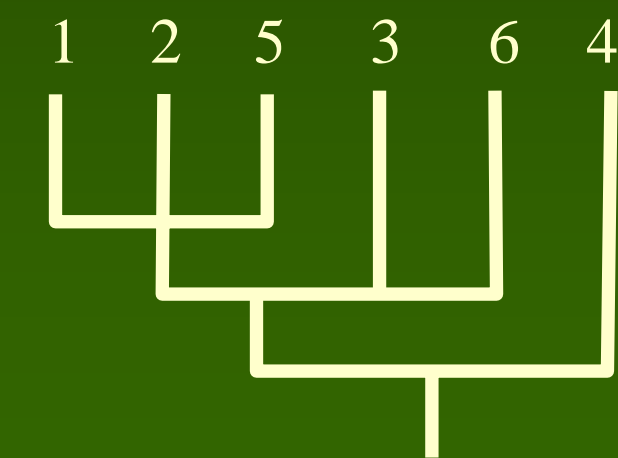
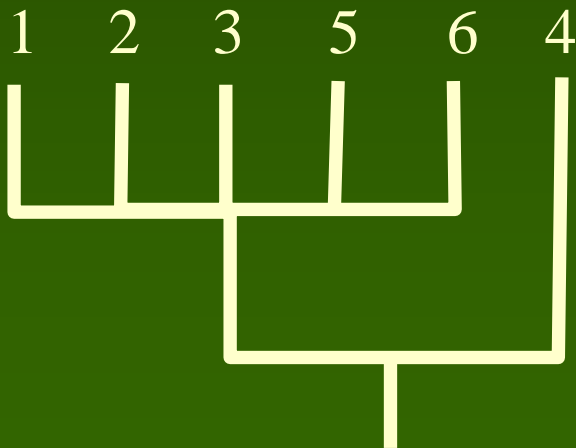
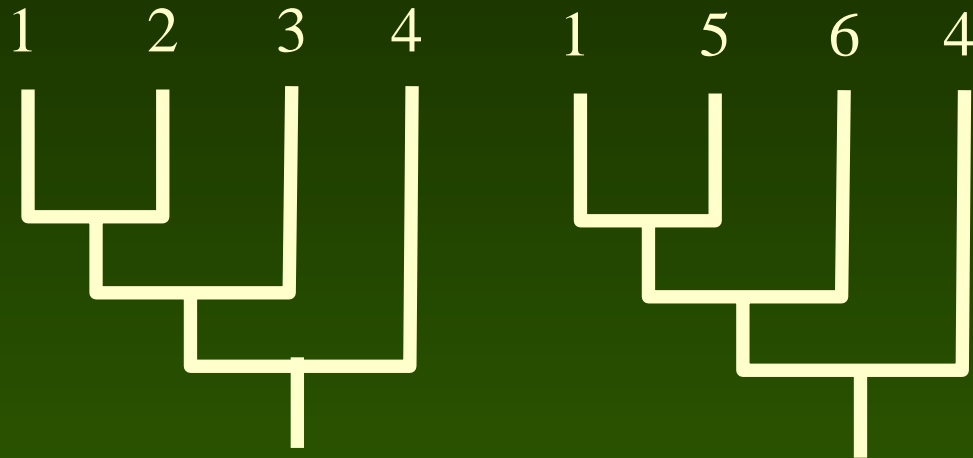


- Build tree using the clusters retained:
(1 2 3 5 6)
(1 2 5)

MinCut algorithm

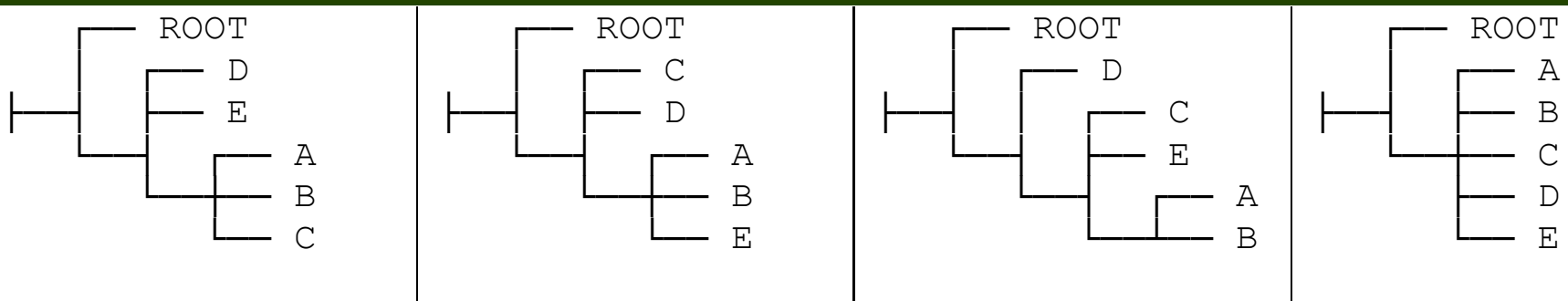
- Advantages:
 - Quick
 - Retains elements common to all input trees
- Disadvantages
 - Can introduce clades not present in any source tree
- Unexpected behaviour when trees differ in size
 - Modification by Page:
 - When cutting, try to save the “uncontradicted” relations at the cost of “contradicted” ones

MRP vs MinCut algorithm



MinCut: problem

- MinCut algorithm can also produce clades that are not present in any source tree



Source trees

MinCut

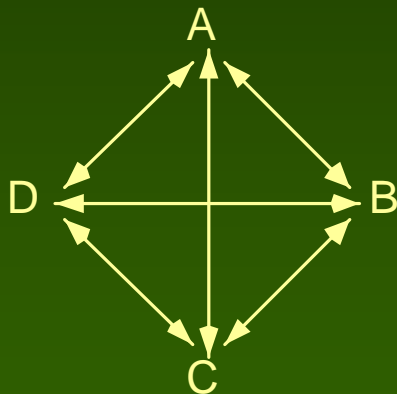
MRP

Step Matrix method (Turner, unpublished)

- Treat each cladogram as a highly ordered single character
- Analyse these characters in a standard parsimony analysis
- Possible uses
 - Combining host-parasite cladograms
 - Combining gene trees
- Advantages
 - Flexibility
- Disadvantages
 - Limited nr of taxa can be dealt with
 - Not available as computer program

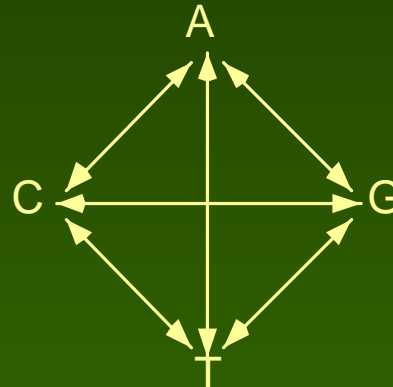
Step Matrix method

- Use step matrix to specify the steps for each character transition



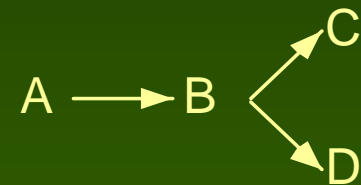
	A	B	C	D
A	.	1	1	1
B	1	.	1	1
C	1	1	.	1
D	1	1	1	.

(a) unordered



	A	C	G	T
A	.	5	5	1
C	5	.	1	5
G	5	1	.	5
T	1	5	5	.

(b) unordered,
differential weighting

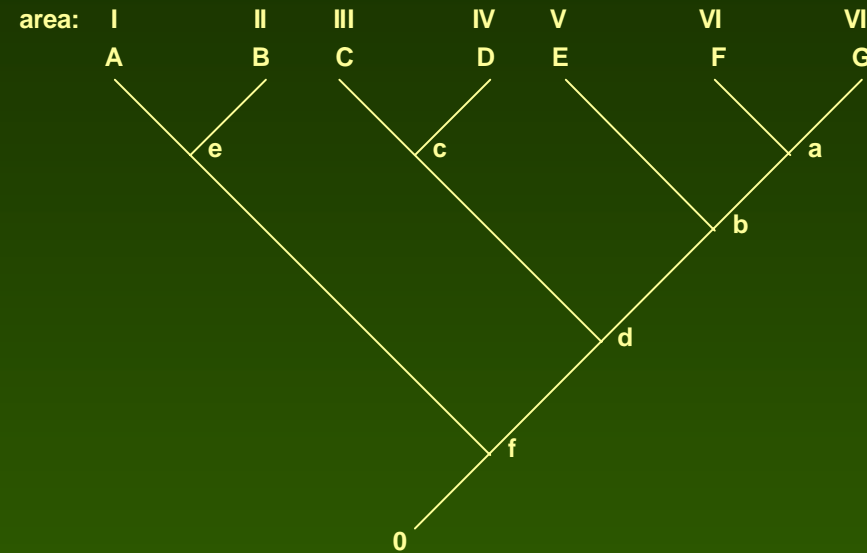


	A	B	C	D
A	.	1	2	2
B	i	.	1	1
C	i	i	.	i
D	i	i	i	.

(e) branched ordered

Step Matrix method

- Apply this method to cladograms as complex branched character state trees



	0	A	B	C	D	E	F	G	a	b	c	d	e	f	
0	0	89	89	89	89	89	89	89	89	89	89	89	89	8	9
A	10	0	99	99	99	99	99	99	99	99	99	99	i	i	
B	10	99	0	99	99	99	99	99	99	99	99	99	i	i	
C	10	99	99	0	99	99	99	99	99	i	i	99	99	i	
D	10	99	99	99	0	99	99	99	99	i	i	99	99	i	
E	10	99	99	99	99	0	99	99	99	i	99	i	99	i	
F	10	99	99	99	99	99	0	99	i	i	99	i	99	i	
G	10	99	99	99	99	99	99	0	i	i	99	i	99	i	
a	10	99	99	99	99	99	10	10	0	i	99	i	99	i	
b	10	99	99	99	99	10	20	20	10	0	99	i	99	i	
c	10	99	99	10	10	99	99	99	99	0	i	99	99	i	
d	10	99	99	20	20	20	30	30	20	10	10	0	99	i	
e	10	10	10	99	99	99	99	99	99	99	99	99	0	i	
f	10	20	20	30	30	30	40	40	30	20	20	10	10	0	

```
begin assumptions;
```

```
  usertype GS = 18
```

Step Matrix method

- Input this table as a step-matrix character for PAUP

```
      C   A   E   C   D   E   F   G   H   I   j   k   l   m   n   o   p   q
[0]   .   M   M   M   M   M   M   M   M   M   M   M   M   M   M   M   M   M
[A]   X   .   Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   i   Q   i   i
[B]   X   Q   .   Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   i   Q   i   i
[C]   X   Q   Q   .   Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   i   i   i
[D]   X   Q   Q   Q   .   Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   i   i   i
[E]   X   Q   Q   Q   Q   .   Q   Q   Q   Q   i   i   i   i   Q   Q   Q   i
[F]   X   Q   Q   Q   Q   Q   .   Q   Q   Q   i   i   i   i   Q   Q   Q   i
[G]   X   Q   Q   Q   Q   Q   Q   .   Q   Q   Q   i   i   i   Q   Q   Q   i
[H]   X   Q   Q   Q   Q   Q   Q   Q   .   Q   Q   Q   i   i   Q   Q   Q   i
[I]   X   Q   Q   Q   Q   Q   Q   Q   Q   .   Q   Q   Q   i   Q   Q   Q   i
[j]   X   Q   Q   Q   Q   Q   10 10  Q   Q   Q   .   i   i   i   Q   Q   Q   i
[k]   X   Q   Q   Q   Q   Q   20 20 10  Q   Q   10  .   i   i   Q   Q   Q   i
[l]   X   Q   Q   Q   Q   Q   30 30 20 10  Q   20 10  .   i   Q   Q   Q   i
[m]   X   Q   Q   Q   Q   Q   40 40 30 20 10 30 20 10  .   Q   Q   Q   i
[n]   X   10 10  Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   .   Q   i   i
[o]   X   Q   Q   10 10  Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   .   i   i
[p]   X   20 20 20 20  Q   Q   Q   Q   Q   Q   Q   Q   Q   Q   10 10  .   i
[q]   X   30 30 30 30 50 50 40 30 20 40 30 20 10 20 20 10  .
```

```
;
```

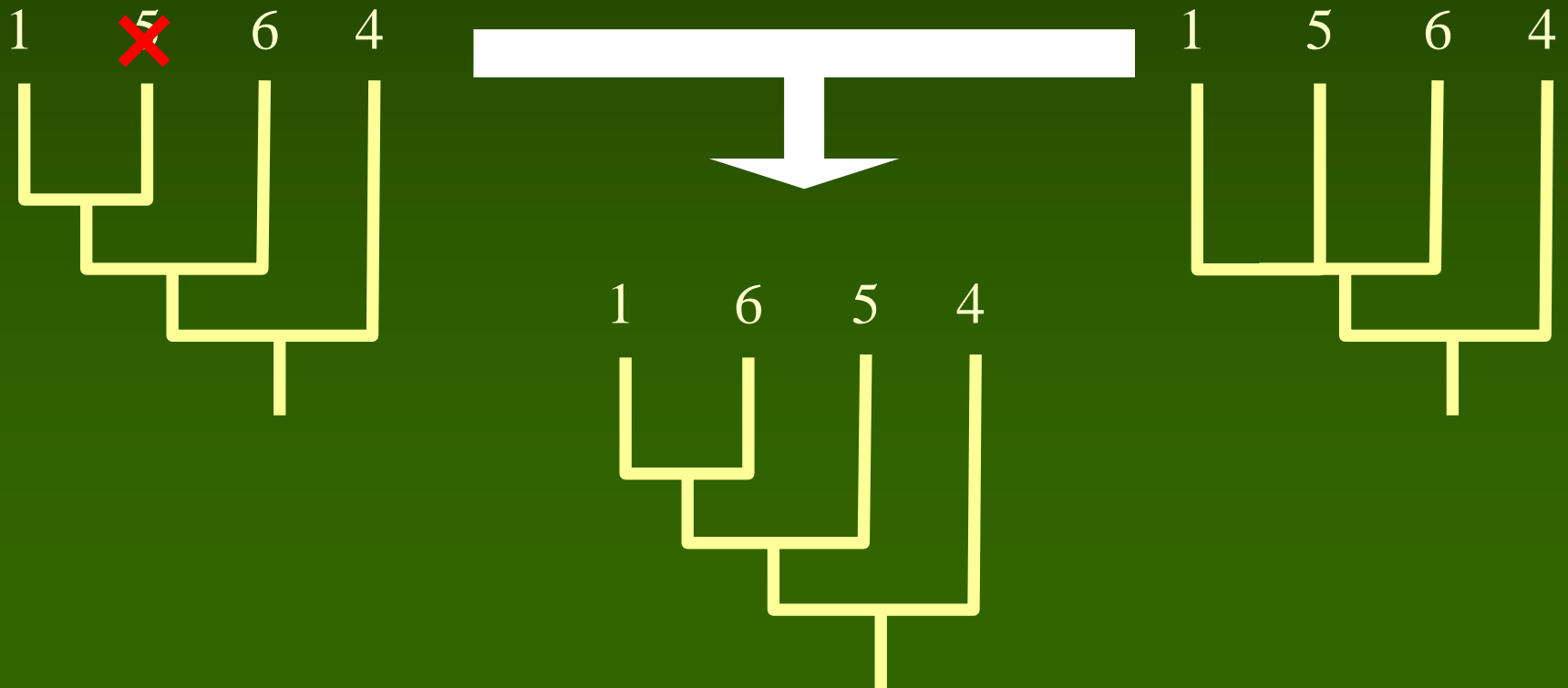
```
  ancstates *allZero = 0:1;
```

```
  typeset *set = GS:1;
```

```
end;
```

General problem: Taxon sampling effects

- Not including some taxa can lead to “clades” that will find their way into the supertree, but are actually absent in the complete source trees



Implementations

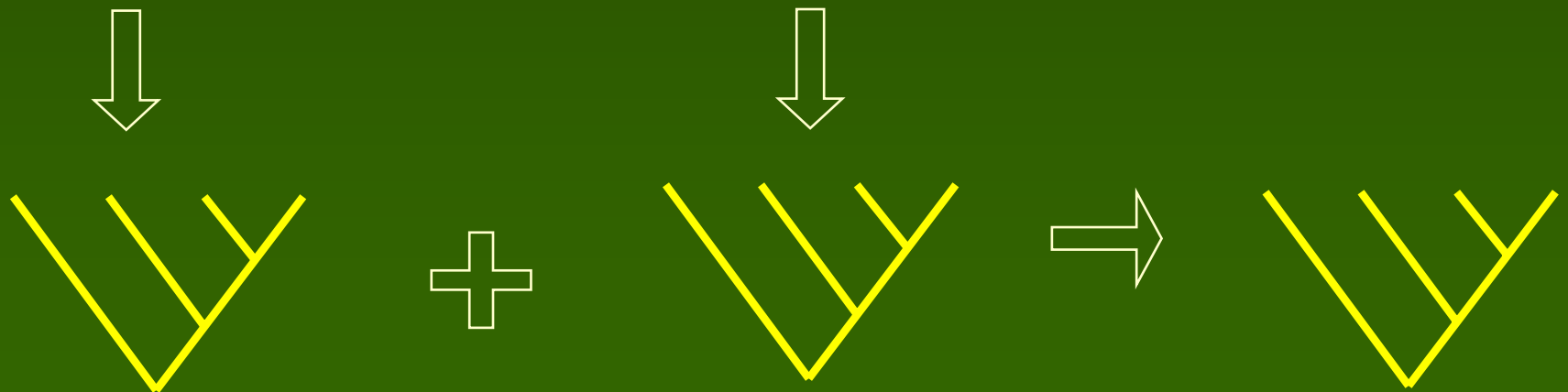
- MRP
 - Parsimony: TNT
 - Semi strict supertrees: TNT
 - (Flip: Rainbow)
- MinCut algorithm (Semple & Steel 2000, modified by Page)
 - Executable: <http://darwin.zoology.gla.ac.uk/~rpage/supertree/>
 - On-line: <http://darwin.zoology.gla.ac.uk/cgi-bin/supertree.pl>
 - (Rainbow): supertree
- Step matrix method
 - Manual implementation in Paup, TNT

Supertrees are just one option for combining data

Combined analysis



Separate analysis



Combined analysis (“Total Evidence”)

- Put all characters together in a datamatrix and analyze them all together.
- General properties
 - All characters are compared individually.
 - Characters from small datasets are not implicitly given a stronger weight

Separate analysis (“Taxonomic congruence”)

- Distinguish functionally or evolutionary different partitions in the data
- Analyse data from different “partitions” separately
- Combine the results
- General properties
 - Characters in one set are not individually compared to those in another set.
 - Large datasets are not automatically outnumbering smaller sets

Combining data?

The problem of combination vs. separate analysis is often reduced to one of the following questions:

- Do they represent the same phylogeny?
- Do they have comparable evolutionary rates?
- Can they be analyzed using the same model?

Why not combine?

- The “Dilution” argument:
“the inclusion of many different characters increases the chance that support for true phylogenetic grouping coming from reliable characters may be diluted by random or systematic errors from unreliable characters” (Bull et al, 1993)
- Is this a valid argument?

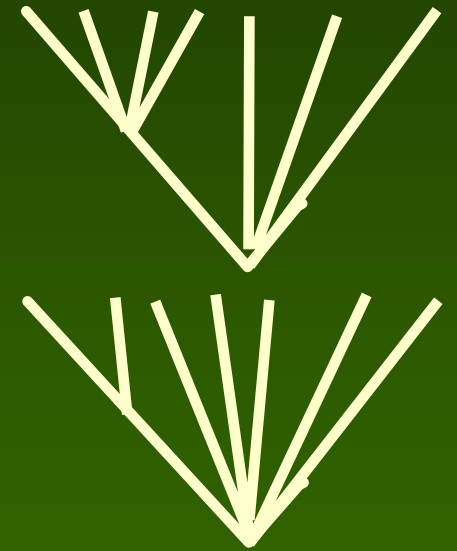
Compromise?

- Find out first if different character sets give different results or not
- Then decide whether to combine or not

Combining data: the basic question

Why deal differently with single characters and entire datasets?

	A	B	C	D	E	F	G	
char.1	+	+	+	+	-	-	-	
Char.2	+	+	-	-	-	-	-	
Char.3	-	-	-	+	+	+	-	?



Combining data: conclusions

- If datasets are not conflicting, they can be combined
- If datasets are conflicting, there is no reason not to combine them
- If you have to analyze different datasets separately
 - Investigate different methods for combination
 - Do not treat each clade in a tree as if it is a real clade - it may be an artifact without any character support.

