

A Phylogenetic Network Construction due to Constrained Recombination

Mohd. Abdul Hai Zahid

Research Scholar

*Research Supervisors: **Dr. R.C. Joshi***

Dr. Ankush Mittal

*Department of Electronics and Computer Engineering,
Indian Institute of technology, Roorkee*

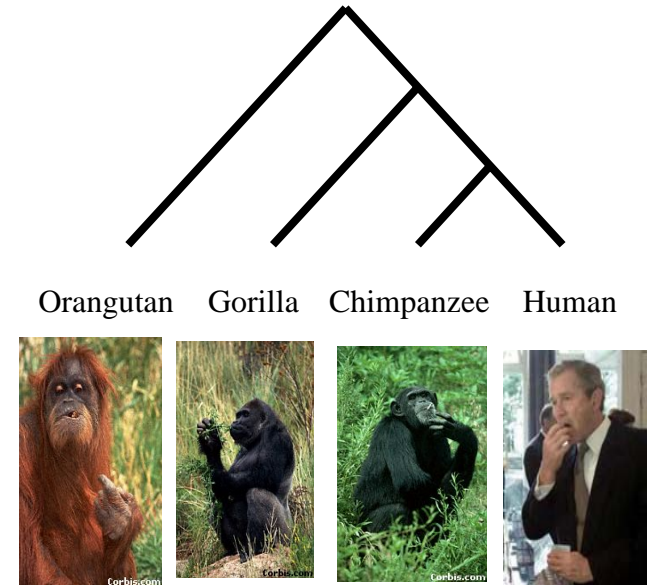
December, 27, 2006

Outline

- ❑ ***Phylogenetics and its applications***
- ❑ *Phylogenetic networks*
 - ❑ *Non-tree like events*
 - ❑ *Existing network reconstruction methods*
 - ❑ *Proposed Algorithm*
- ❑ ***Phylogenetic supertrees***
 - ❑ *Need for supertree methods*
 - ❑ *desirable properties*
 - ❑ *existing methods*
 - ❑ *Proposed Algorithm*

Phylogenetics

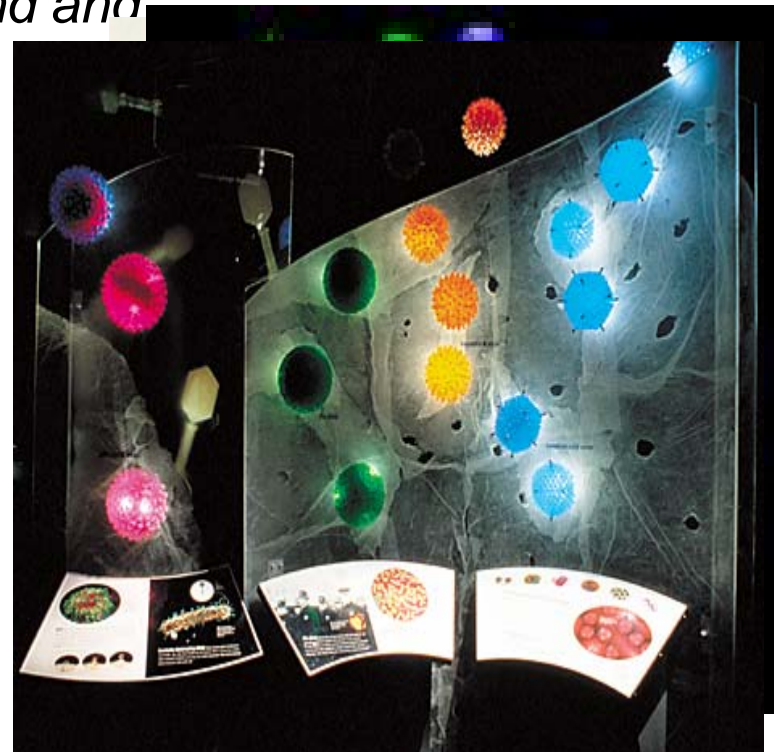
- *phylogenetics is the study of the evolutionary relationships among different species or genes.*
- *Morphological characters were used to classify the species for years.*
- *The increasing availability of DNA and amino acid data has led to the wider interest in computational methods for the tree construction.*
- *The advantage of using genetic data is*
 - *unambiguous characters (A,C,G,T)*
 - *molecular data can be converted to numerical form, which can be used for the computational and mathematical analysis.*



(From university of Arizona)

Applications

- *Big genome sequencing projects just produce the data, which is meaningless until analyzed and classified properly.*
- *Evolutionary history relates all organisms and genes, and helps to understand and predict:*
 - *Interaction between genes*
 - *Drug design*
 - *Predicting functions of genes*
 - *Origins and spread of disease*
 - *Origins and migrations of humans*



Phylogenetic networks

Phylogenetic tree construction methods *failed to find true relationship* between the species not because of *wrong genes were selected or methods are not adequate to do so.*

It is because of the events

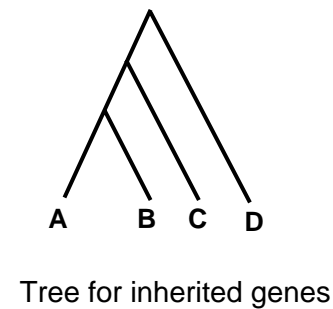
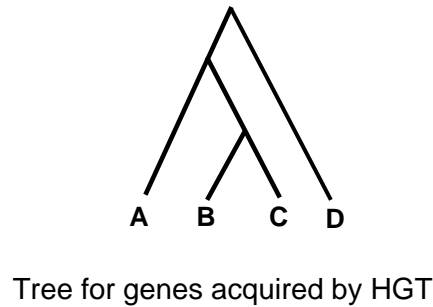
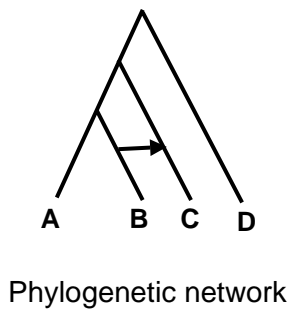
- *Horizontal gene transfer*
- *Hybridization*
- *Homoplasy*
- *Genetic recombination*

SNEATH, P. H. A.. “Cladistic representation of reticulate evolution.” Syst. Zool. 24:360–368. 1975

Non-tree like events

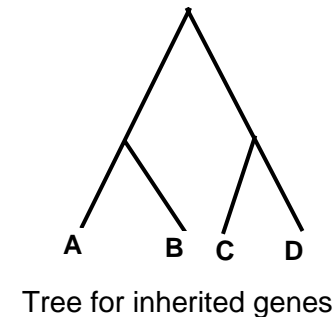
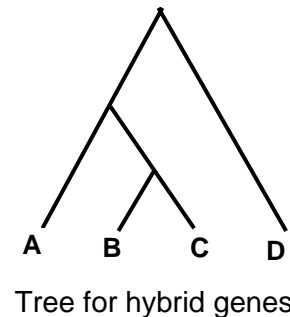
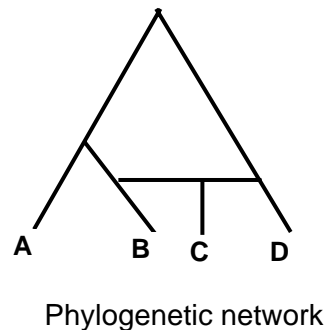
Horizontal gene transfer

Between similar species, through asexual process



Hybridization

Between different species, through sexual process



Non-tree like events cont...

Homoplasy

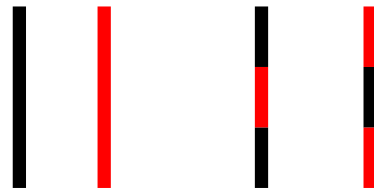
Similarity despite of separate ancestry (evolutionary noise)

P.	A	C	A	T	A	C	G
Q.	G	T	A	T	A	C	G
R.	G	G	A	C	A	T	G
S.	G	C	A	C	A	C	A

Homoplasy is change of state twice or more Ex. Site 2.

Genetic recombination

Within the same lineages



Homologous chromosomes exchanging genetic material

Phylogenetic Networks under Constrained Recombination

Binary sequences (SNP)

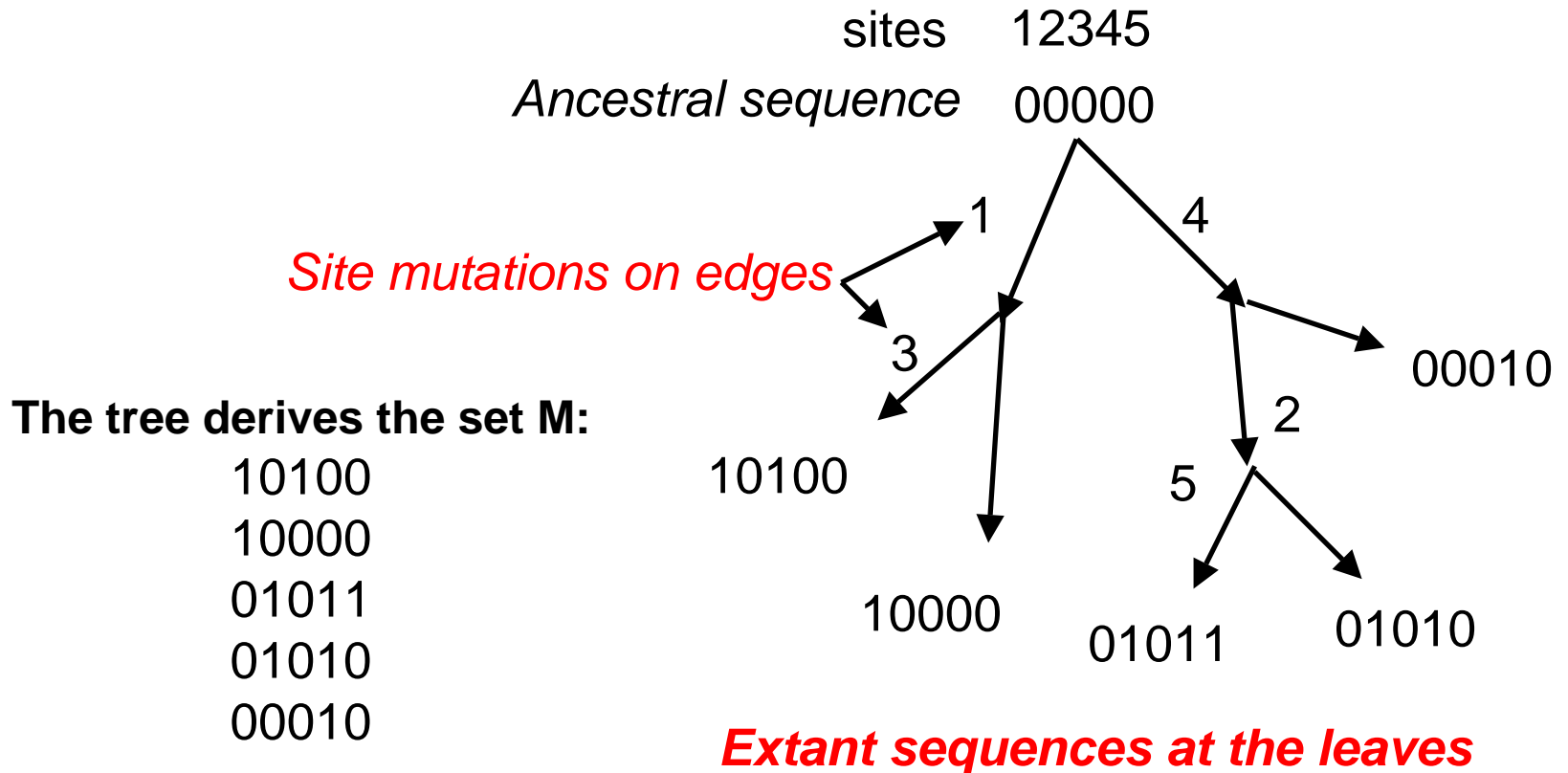
A Single Nucleotide Polymorphism (SNP) (pronounced "snip"), is a small genetic change, or variation, that can occur within a specie 's DNA sequence.

Example AAGGTTA ---ATGGTTA.

Each site in the sequence represents a SNP (single nucleotide polymorphism), a site where two of the four possible nucleotides appear in the population with the frequency above some threshold

The order of the sites is fixed (on a chromosome) in SNPs. In contrast to a set of taxonomic characters, where the given order is arbitrary.

The Perfect Phylogeny Model for binary sequences



The Perfect Phylogeny Problem

Given a set of sequences M we want to find, if possible, a perfect phylogeny that derives M . Remember that each site can change state from 0 to 1 only once.

n will denote the number of sequences in M , and m will denote the length of each sequence in M .

		m
		01101001
n		11100101
		10101011

The sequence matrix M

The 4-Gamete Test

This will test whether the given sequences derive a perfect phylogenetic tree or not.

*Arrange the sequences in a matrix. Then (with **no** duplicate columns), the sequences can be generated on a **unique** perfect phylogeny if and only if no two columns (sites) contain all four pairs:*

0,0 and 0,1 and 1,0 and 1,1

An Example of Gamete test

The tree derives the set M:

10100

10000

01011

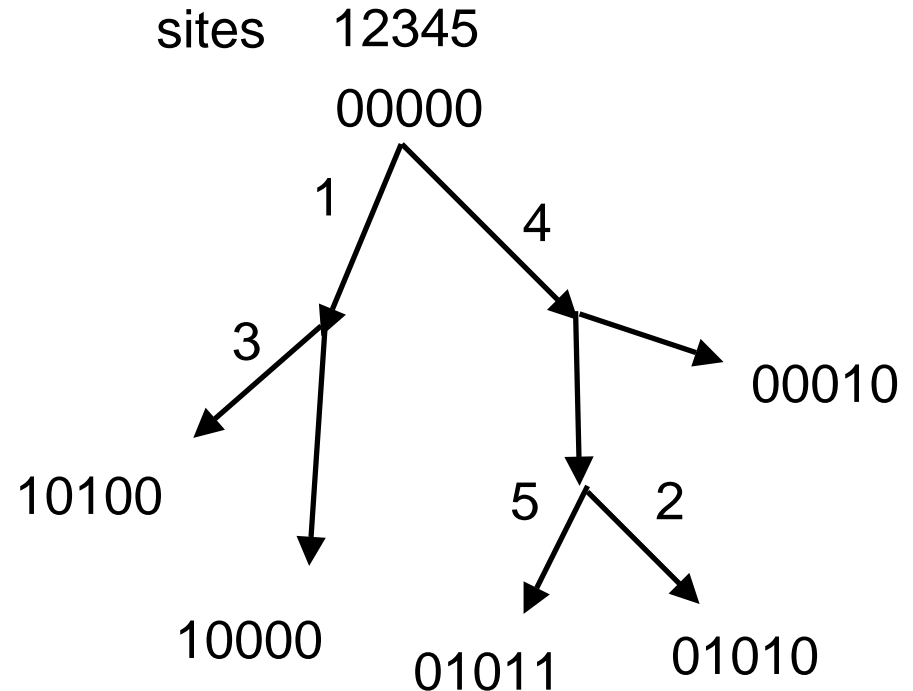
01010

00010

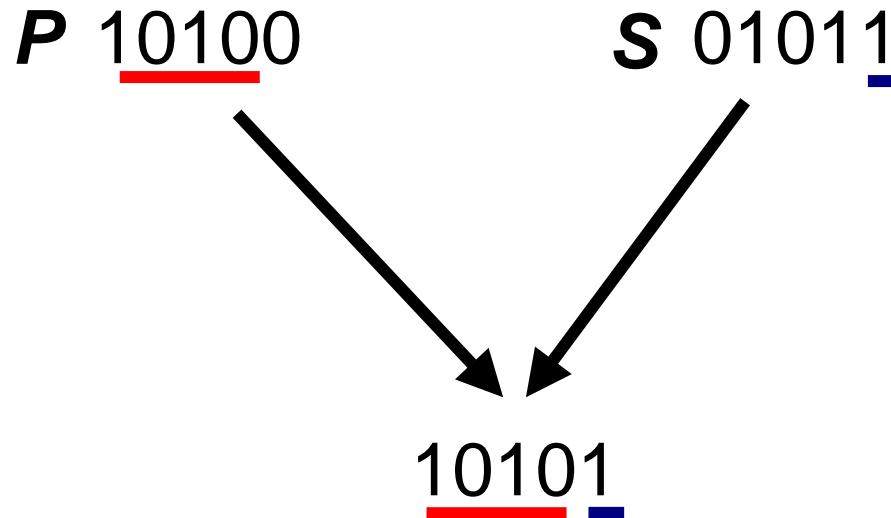
10101 (new)

**Columns 4 and 5 fails
the Gamete test. This is
called a **Conflict**.**

Conflicts are due to **RECOMBINATION and is
often encountered in real sequences**



Recombination in SNP



The first 4 sites come from P (Prefix) and the sites from 5 onward come from S (Suffix).

Recombination Network with now node

The tree derives the set M:

10100

10000

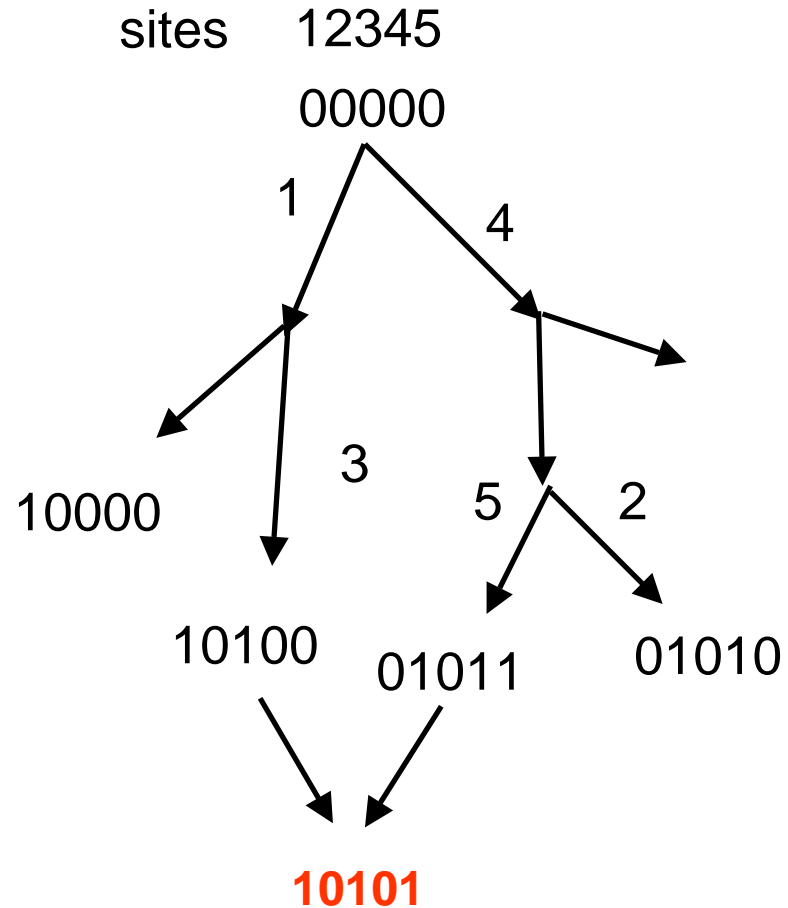
01011

01010

00010

10101 (new)

*Phylogenetic Network for M
with single recombination
event.*



Biologically Significant Network

Given M there may exist many Network derivations for the sequence matrix M .

A biologically significant network is

Perfect Phylogeny

Otherwise a little deviation from tree (Network) with minimum recombination events.

Minimizing number of recombination events

The problem of finding a phylogenetic network that creates a given set of sequences M , and minimizes the number of recombinations, is NP-hard. (Wang et al 2000) (Semple 2004).

*Wang et al. explored the problem of finding a phylogenetic network where the recombination cycles are **required to be node disjoint, if possible.***

They gave a sufficient but not a necessary condition to recognize cases when this is possible. $O(nm + n^4)$ time.

Recombination Cycles

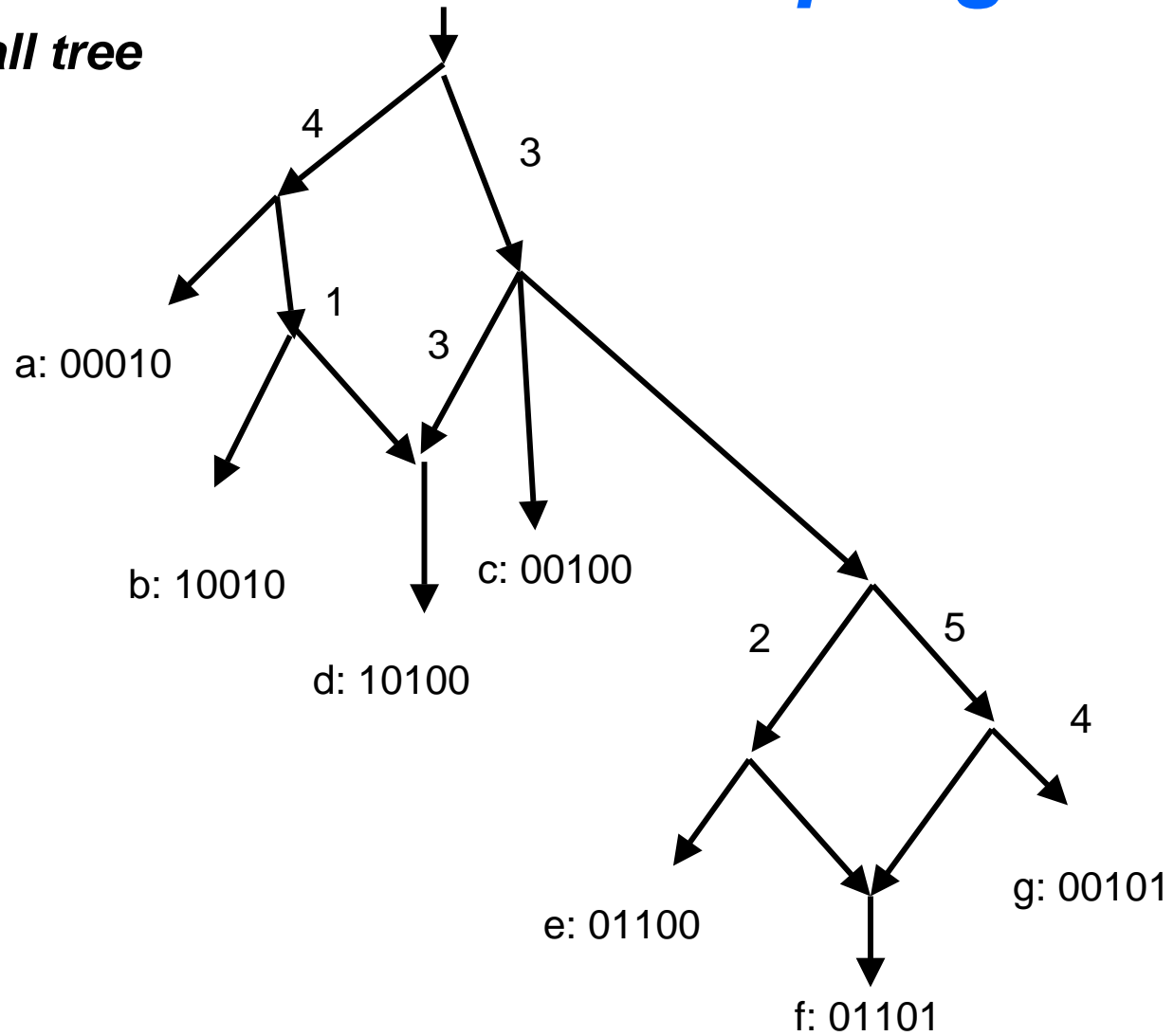
- ***In a Phylogenetic Network, with a recombination node x , if we trace two paths backwards from x , then the paths will eventually meet.***
- ***The cycle specified by those two paths is called a “recombination cycle”.***

Gall trees

- *A recombination cycle in a phylogenetic network is called a “gall” if it shares no node with any other recombination cycle.*
- *A phylogenetic network is called a “galled-tree” if every recombination cycle is a gall.*

Example gall tree

Rooted Gall tree



The Gall Tree Construction Algo.

- 1. Input binary sequences***
- 2. Sort the them in ascending order***
- 3. Find similarity and dissimilarity between the sequences corresponding to '1'.***
- 4. Classify the nodes into predefined classes***
- 5. Make Parent and Child Tables***
- 6. Use the information available from step 3, 4 and 5 to construct the gall tree***

Classification of nodes

We follow the rule based classification approach

The nodes are classified into three classes Null, Mutation, and Recombination

The nodes with least binary sequence values and no similarity between them, belongs to NULL class

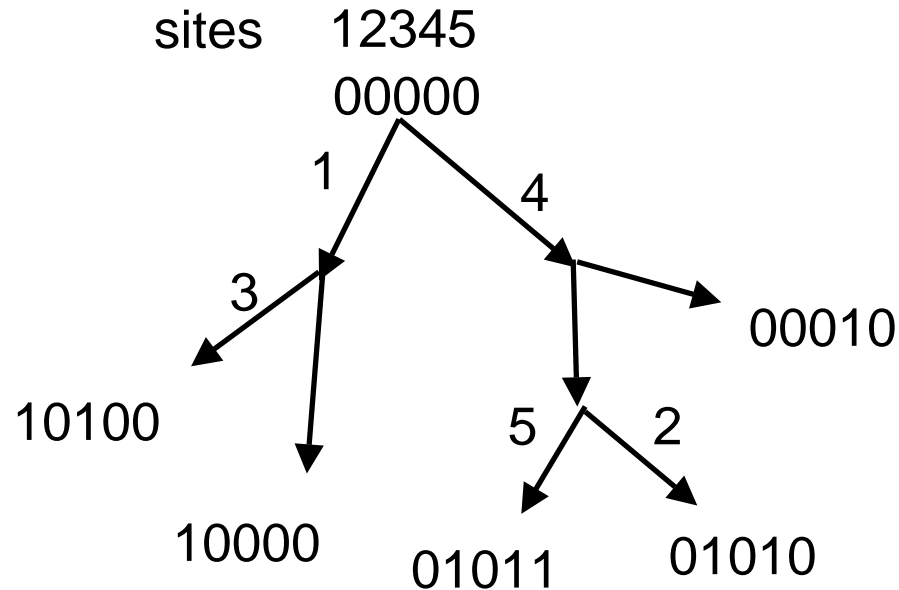
The nodes with similarity to a single node are classified as MUTATION class

The nodes with similarity to a more than one node are classified as RECOMBINATION class

Mathematical Basis for rule formation

Lemma 1

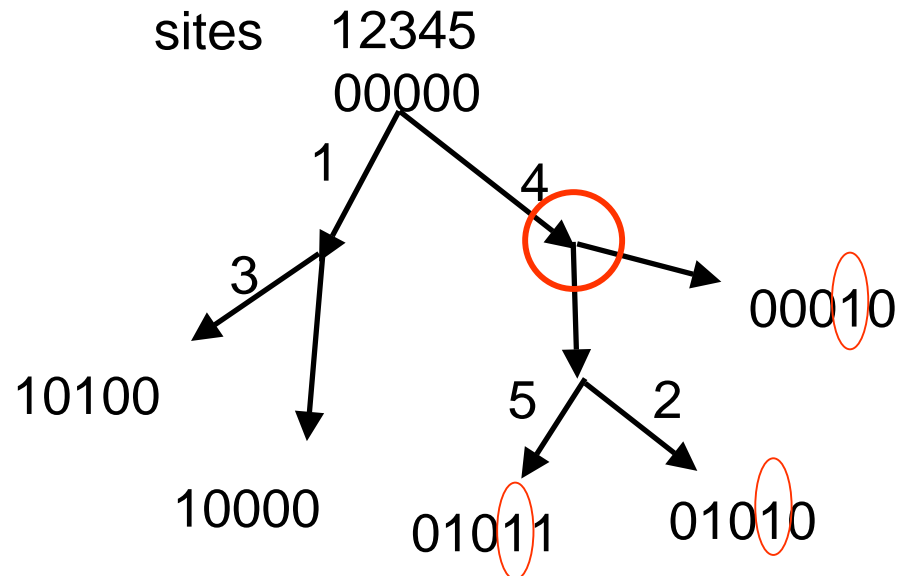
If a node v' is the result of mutation from its parent v then $v < v'$, when the sequences are considered as the binary numbers.



Mathematical Basis for rule formation

Lemma 2

Let S and S' be the sequences of the children of node V . if S' is not the result of the mutation or recombination in S then the similarity between S and S' is due to common ancestry.

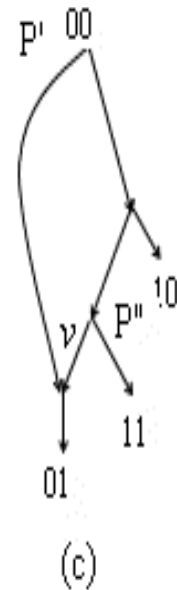
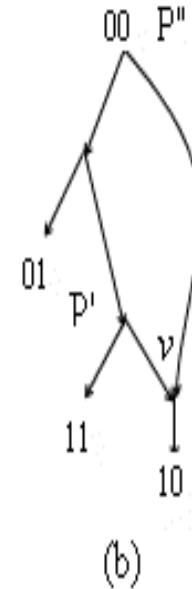
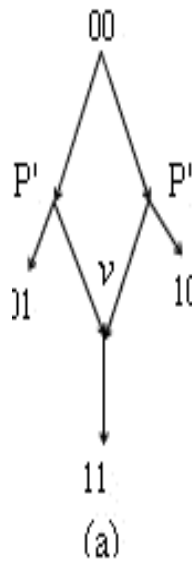


Mathematical Basis for rule formation

Lemma 3

Let v be a recombination node with sequence S . if P' and P'' are two parent nodes of v , with sequences S' and S'' respectively, then any of the following should hold.

- $S' > S$ and $S'' > S$**
- $S' > S$ and $S'' < S$**
- $S' < S$ and $S'' > S$**



An Example

Input sequences:

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

*Similarity and dissimilarity matrix:
(sim,dis)*

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

Example contd...

Classification of Nodes

Node Label	Type	Count
A	Null	0
B	Mutation	1
C	Null	0
D	Recombination	2
E	Mutation	1
F	Recombination	3
G	Mutation	1

Child Table

Node Label	Child List
A	B
B	D
C	D,E,F,G
D	Null
E	F
F	Null
G	F

Parent Table

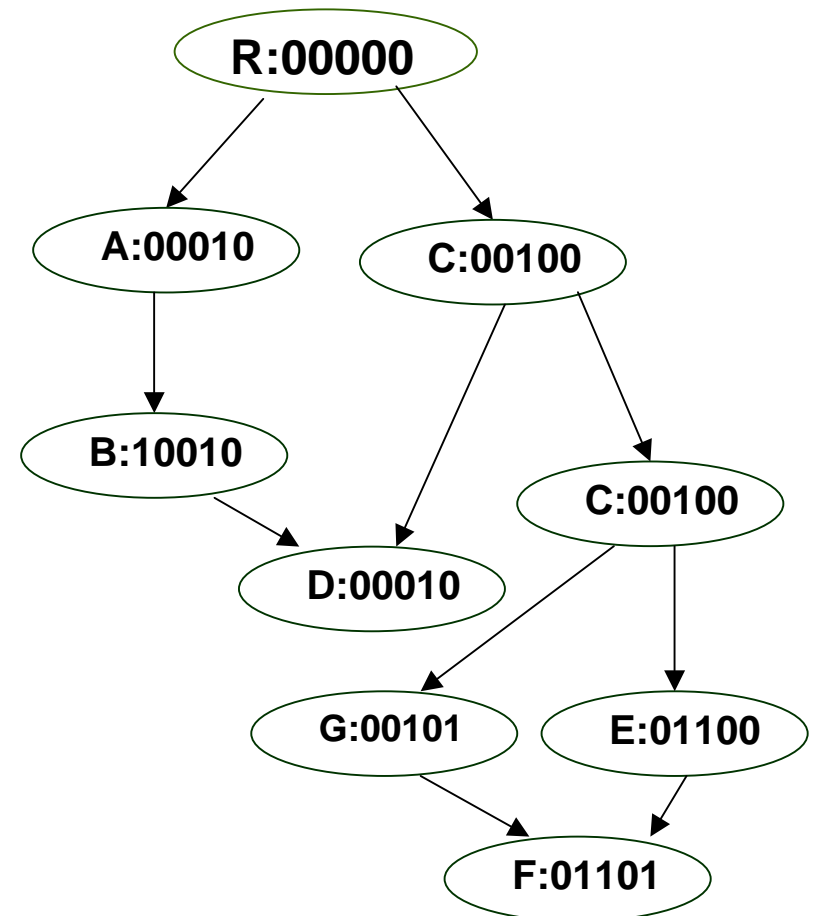
Node Label	Parent List
A	Null
B	A
C	Root
D	B,C
E	C
F	E,G
G	C

Research problems *cont...*

4. Phylogenetic networks :recombination – an example

Node I	Type	Count
A	Null	0
B	Mutation	1
C	Null	0
D	Recombination	2
E	Mutation	1
F	Recombination	3
G	Mutation	1

Node Label	Child List
A	B
B	D
C	D,E,F,G
D	Null
E	F
F	Null
G	F



Contribution of the work

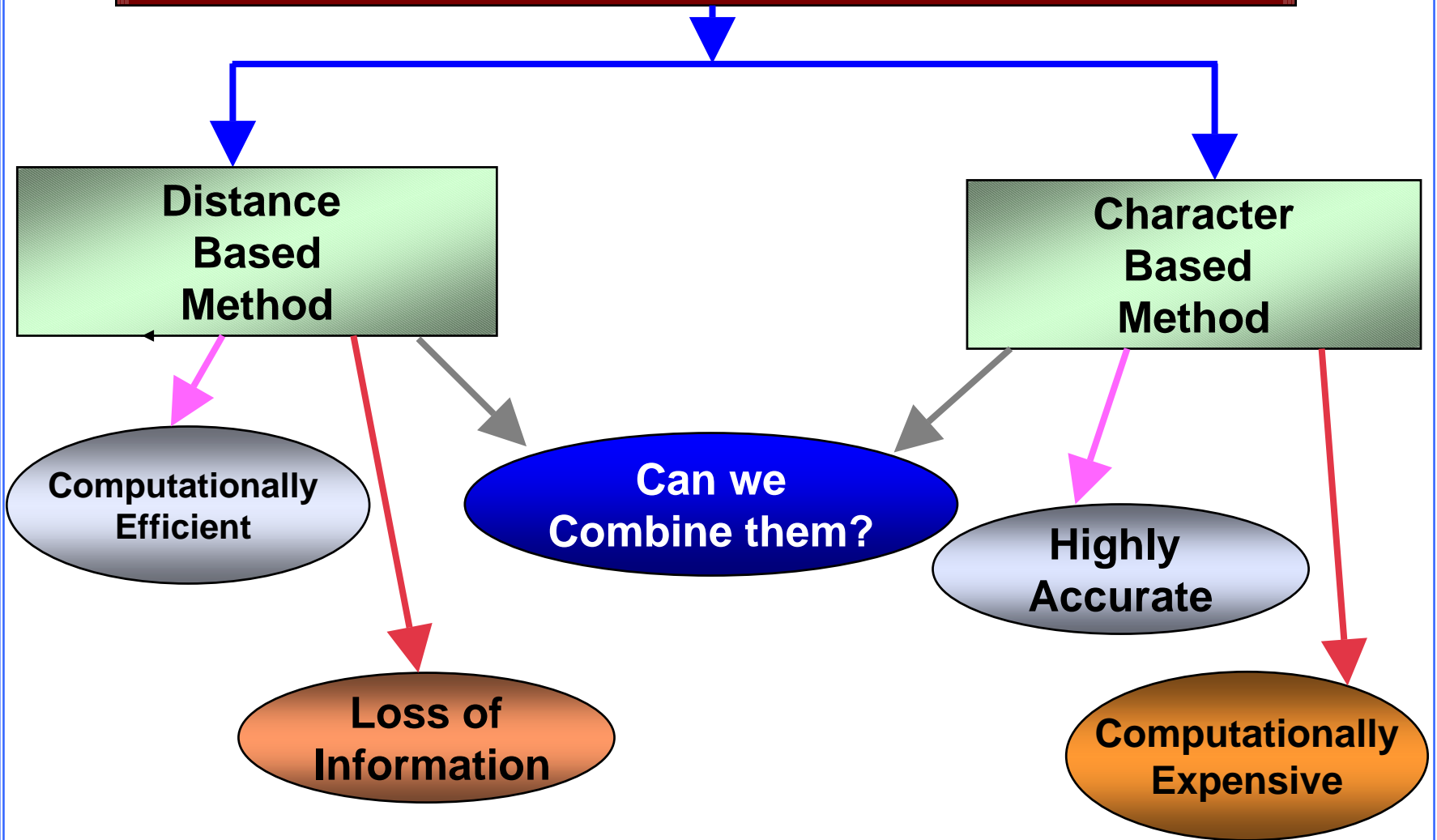
- 1. Used similarity and dissimilarity to avoid information loss***
- 2. Taken a row based search instead of column based search for detecting recombination. This reduces the time complexity of the algorithm.***
- 3. Given better network visualization that is close real visualization and biologically significant.***

- ✓ ***Phylogenetics and its applications***
- ✓ *Phylogenetic networks*
 - ✓ *Non-tree like events*
 - ✓ *Existing network reconstruction methods*
 - ✓ *Proposed Algorithm*
- ***Phylogenetic supertrees***
 - *Need for supertree methods*
 - *desirable properties*
 - *existing methods*
 - *Proposed Algorithm*

Phylogenetic supertrees

- *There are nearly 1.7 million described species.*
- *Not possible for a single researcher or a small group to construct the "Tree of Life".*
- *No algorithm exist for the construction of the Tree of Life.*
- *Phylogenetic supertree methods combine smaller phylogenetic trees into a single tree in such a way that no branching information carried small trees is lost.*

Phylogenetic tree construction methods



Desirable properties

There exist very few algorithms which satisfy the following properties:

- *The supertree can be computed in polynomial time.*
- *The algorithm preserves the branch information shared by all the input trees.*
- *Changing the order of the input tree should not change the resulting supertree.*
- *Renaming or relabeling leads to the change in the corresponding label in the resulting tree and should not effect the supertree topology.*
- *In case all the input trees are compatible the algorithm should return a tree which displays all the input trees.*

Existing supertree methods

- *If the input trees classifies the same set of leaf nodes, the result of amalgamating them is called Consensus tree (constrained supertree).*
- *Many supertree methods use the consensus methods as the basis for supertree construction.*

Popular consensus methods:

Strict, majority consensus method, loose consensus, Adam's consensus methods, and BUILD by Aho et. al.

Popular supertree methods:

Matrix representation of parsimony (MRP), Mincut, Modified Mincut, RankedTree, MinFlip supertree.

- *BUILD is used as the basis for the supertree construction in Mincut, Modified Mincut and RankedTree.*

Existing supertree methods *cont...*

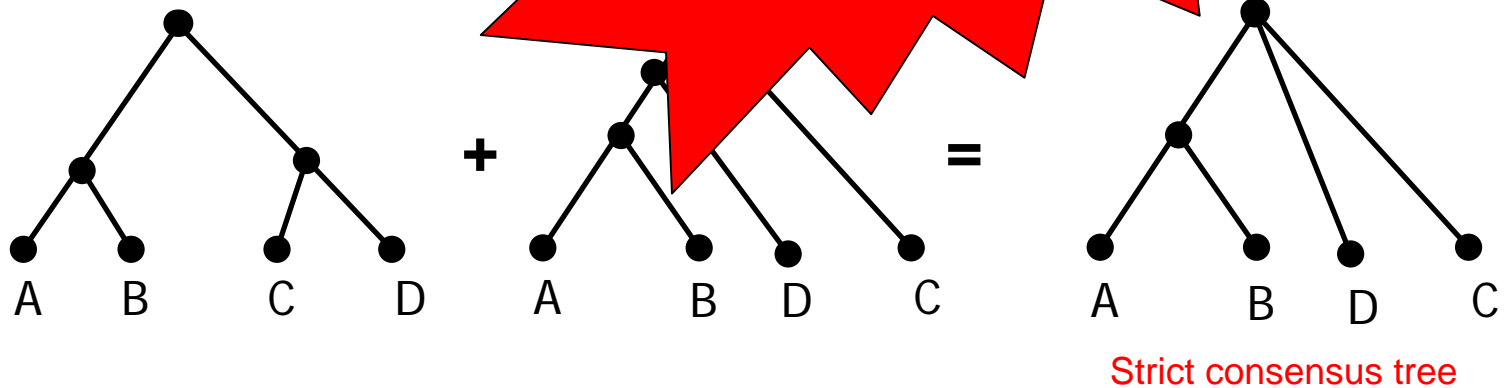
Consensus methods:

Strict consensus methods:

Given a collection of phylogenetic trees, the strict consensus tree is the tree that contains only the clusters common to all the trees.

Majority consensus methods:

Given a collection of phylogenetic trees, the majority consensus tree is the tree that contains only the clusters found in more than half of the trees.



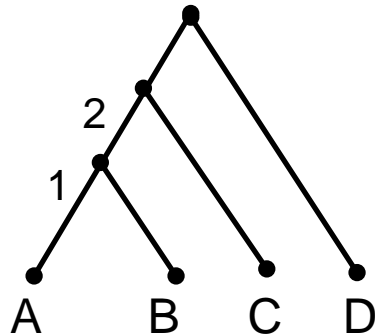
Existing supertree methods *cont...*

Supertree methods:

Matrix representation of parsimony (MRP)

- As a first step, the input phylogenetic tree is converted to a binary matrix.
- In the second step parsimony trees are constructed from the matrix to construct the supertree.
- Exponential time complexity.

Computationally expensive



	1	2
A	0	0
B	1	0
C	1	1
D	1	1

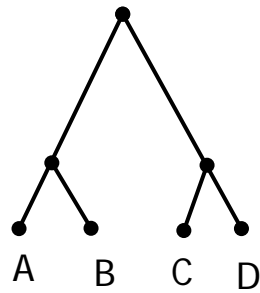
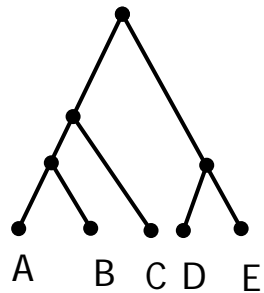
A phylogenetic tree and its binary character matrix.

Existing supertree methods *cont...*

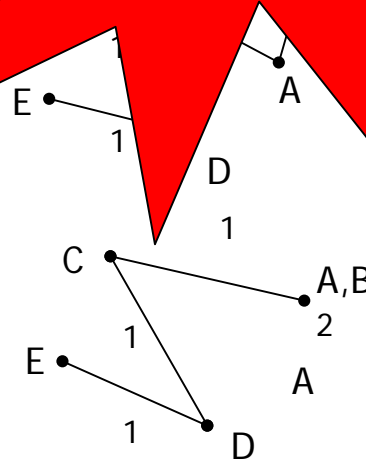
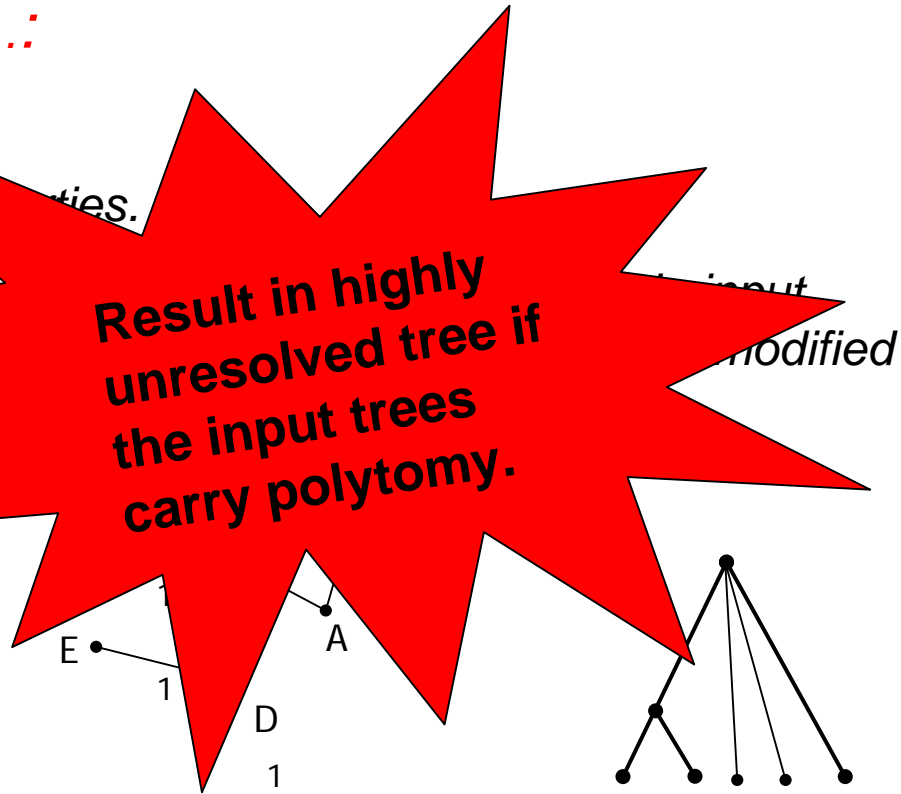
Supertree methods *cont....*:

Mincut:

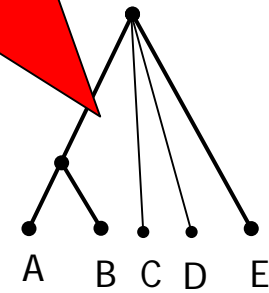
- Satisfy all the desirable properties.
- Extension of BUILD, BUILD* to handle polytomy in input trees. Mincut removes the minimum weight edges from the modified graph to construct supertree.



Input trees



Modified graphs



Supertree

Outline

- ✓ **Phylogenetics and its applications**
- ✓ *Phylogenetic networks*
 - ✓ *Non-tree like events*
 - ✓ *Existing network reconstruction methods*
- ✓ **Phylogenetic supertrees**
 - ✓ *Need for supertree methods*
 - ✓ *desirable properties*
 - ✓ *existing methods*
- **Supertree with ancestral divergence time**
- **Supertree for semi-labeled taxa**
- **Phylogenetic Question Answering System**
- **Conclusion and future work**

Research problems and our approach

Research problems:

- 1. Phylogenetic supertrees.*
- 2. Phylogenetic supertree methods for ancestral time divergence data.*
- 3. Phylogenetic supertree methods for semi-labeled trees.*
- 4. Phylogenetic networks due to recombination.*
- 5. Phylogenetic Question Answering System*

Research problems *cont...*

1. Phylogenetic supertrees

- *Due to the drawbacks of character and distance based methods new techniques to be developed for the construction of the supertrees.*
- *The idea is to combine small phylogenetic trees into a single tree in such a way that the branching information carried by each tree is persevered. (branching information includes clusters, triplets, quartets, and splits).*
- *To construct the tree of life all the desirable properties should be satisfied.*
- *If the input trees shows incompatibility, minimum information can be removed from the input trees to result in a supertree.*

Research problems *cont...*

Phylogenetic supertrees: our approach

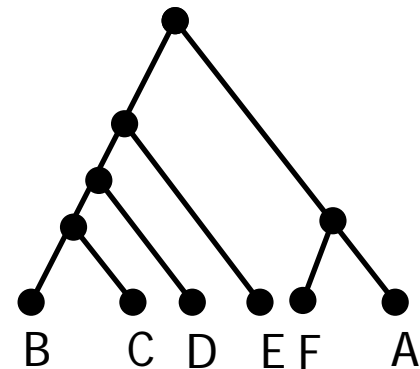
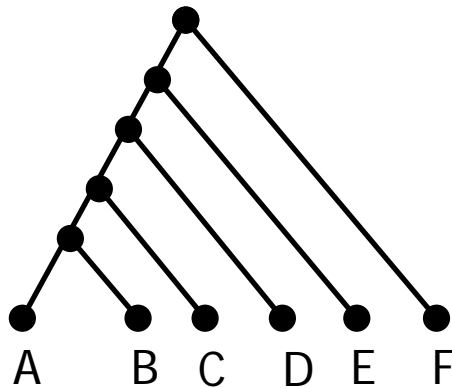
- *We took the recoding based approach.*
- *Used UPGMA variant for the consensus tree construction.*
- *Minimize the least square error to deal with incompatibility.*

Steps in our method:

- 1. Find the restricted trees for common nodes*
- 2. recode the restricted trees by considering path length between the taxa as distance between them.*
- 3. Use UPGMA variant to construct the consensus tree.*
- 4. Add the distinct taxa to the consensus tree in way to minimize the least square error.*

Research problems *cont...*

DISTSUPERTREE: An example



Average distance matrix for the resulting trees

Species	A	B	C	D	E
B	4				
C	4.5	2.5			
D	4.5	3.5	3		
E	4.5	4.5	4	3	
F	4	6	5.5	4.5	3.5

Research problems *cont...*

DISTSUPERTREE: An example cont...

Modified matrix after grouping the taxa B and C

Species	A	BC	D	E
BC	4.25			
D	4.5	3.25		
E	4.5	4.25	<u>3</u>	
F	4	5.75	4.5	3.5

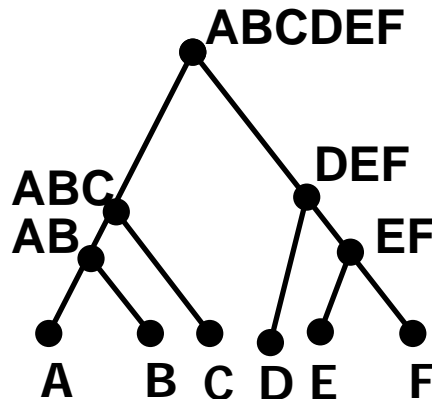
**Leads to wrong conclusion
Ancestor of BC is D not E**

Research problems *cont...*

DISTSUPERTREE: An example cont...

The problem of distance matrix modification can be solved using the distances between the ancestors of the groups and taxa.

Distance Between (AB) and (E)



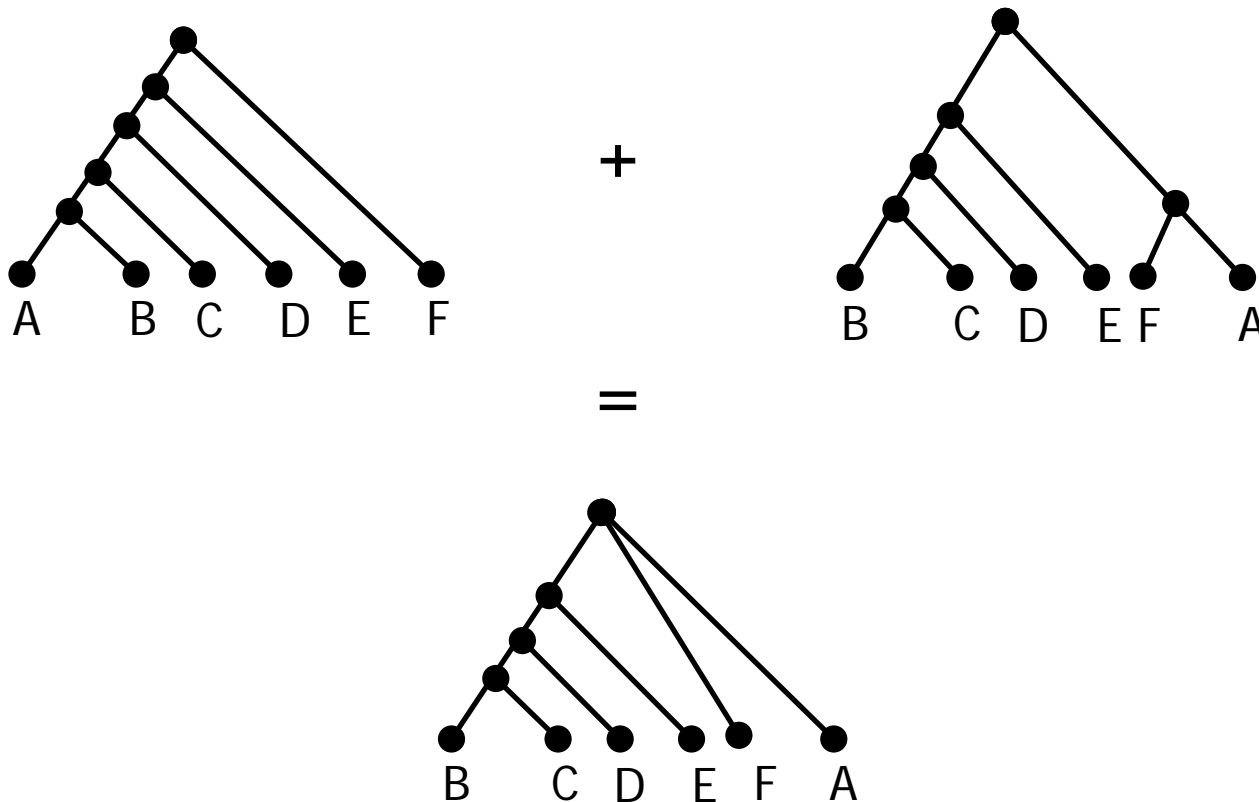
1. Find all the clusters with $X=(AB)$ and $Y=(E)$

$$X = \{(AB), (ABC), (ABCDEF)\}$$

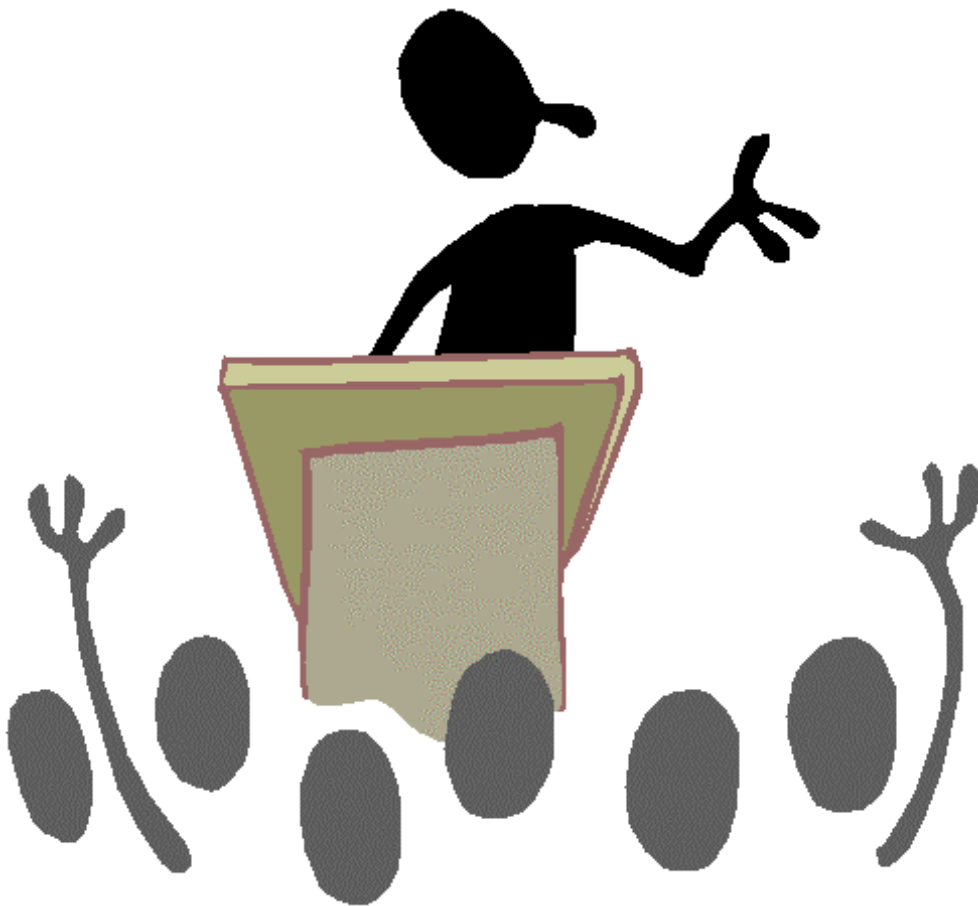
$$Y = \{(E), (EF), (DEF), (ABCDEF)\}$$

Research problems *cont...*

DISTSUPERTREE: The final result



Questions and Answers ...



Thank You ...