Statistically based postprocessing of phylogenetic analysis by clustering

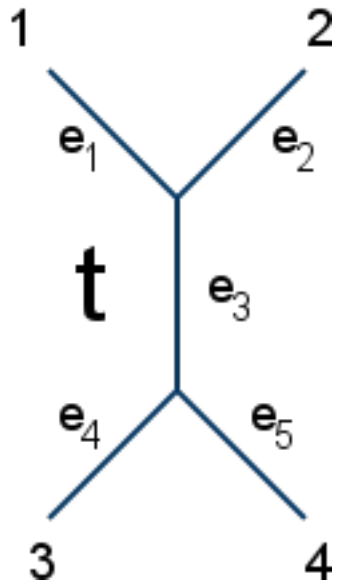
Authors: Cara Stockham, Li-San Wang and Tandy Warnow

David Haws. UKY Sep 8th

Phylogenetic Analysis

1. Collect data (e.g. DNA sequences)
2. Apply tree reconstruction method (e.g. Maximum likelihood, Maximum Parsimony, Mr. Bayes)
3. Now have many equally likely trees. Compute the consensus tree to resolve conflicts. (Biologist typically assume the true tree is among the trees obtained here).

Towards defining Consensus tree...



Every edge(branch) of a tree t induces a split of the leaves {1,2,3,4}.

E.g. e_3 induces {1,2} | {3,4}

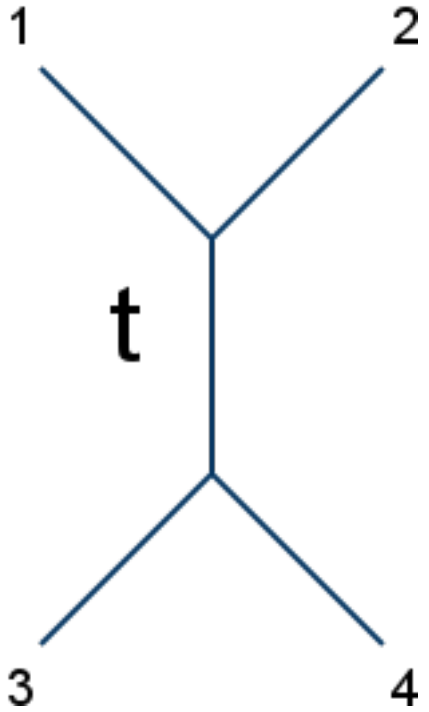
e_4 induces {1} | {2,3,4}

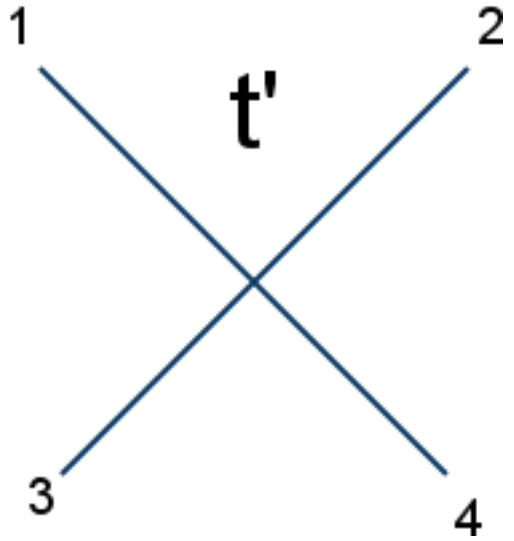
Let E(t) := edge set of tree t.

(Or equivalently the set of splits of t)

Tree t refines tree t' if E(t') is contained in E(t).

Ex.





E(t) ={ {12|34},{1|234},{2|134},{3|124},{4|123} }

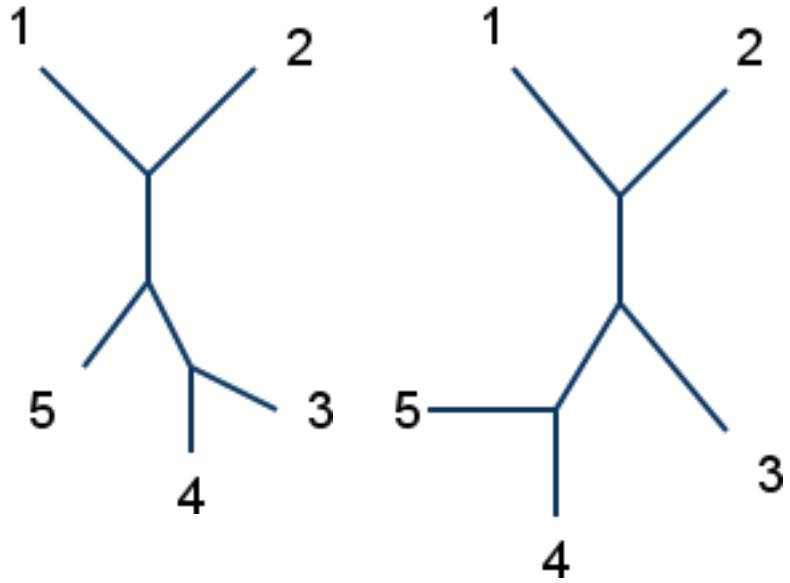
E(t')={ {1|234},{2|134},{3|124},{4|123}}

Note: If t is binary, then it will have n-3 edges. (n = number of taxa/leaves)

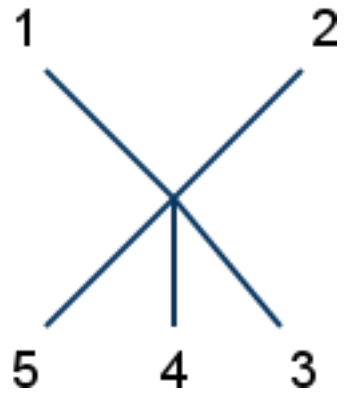
Consensus Tree

The **strict consensus** tree, SC(**T**), of a set of trees **T** is the tree whose edges(splits) are in every tree of **T**.

T =



SC(**T**) =



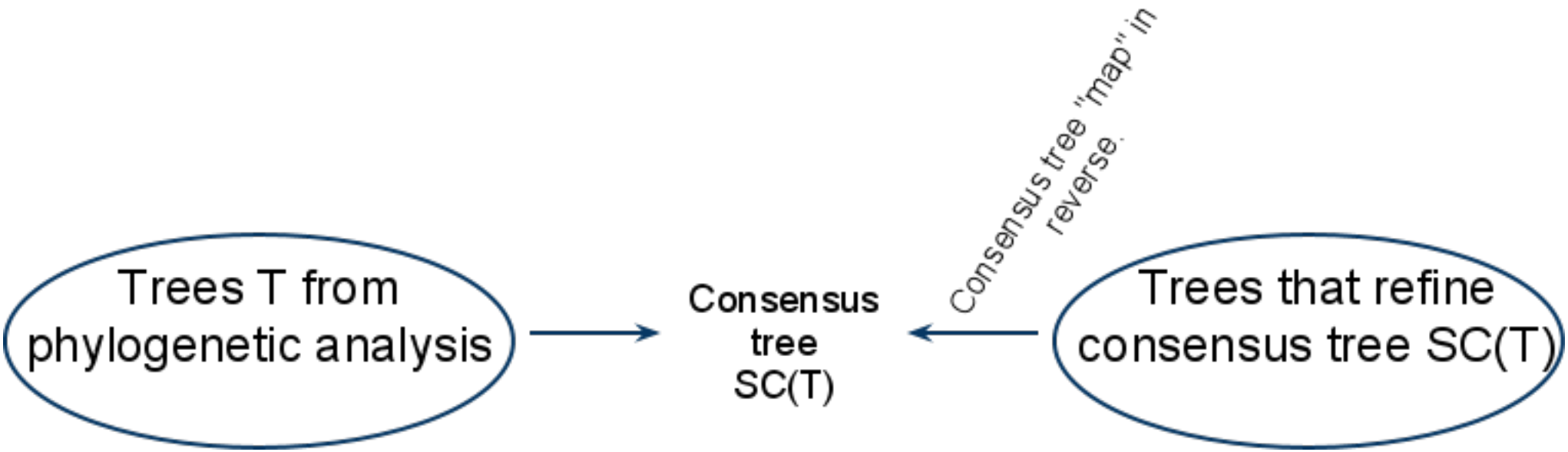
The strict consensus tree (or just consensus tree) is a conservative hypothesis about the true phylogeny.

Note: Other types of "consensus" trees exist, but are not discussed here.

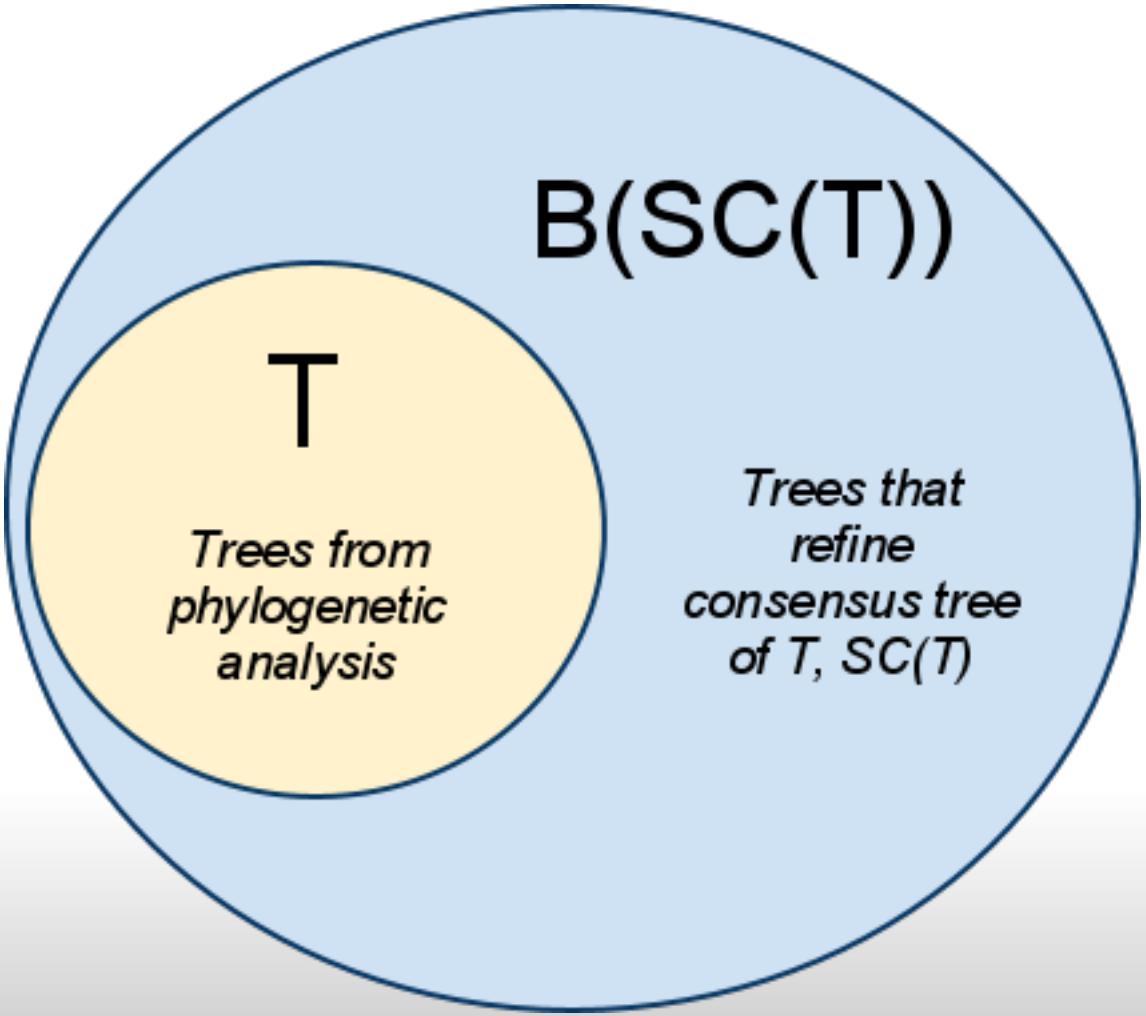
"Because the number of trees (computed) can be overwhelming, biologists replace them with the strict consensus tree, and the original trees are then ignored. Knowing only that the true tree refines this consensus tree."

--C. Stockham et al.

What about all trees that refine the consensus tree SC(T)?

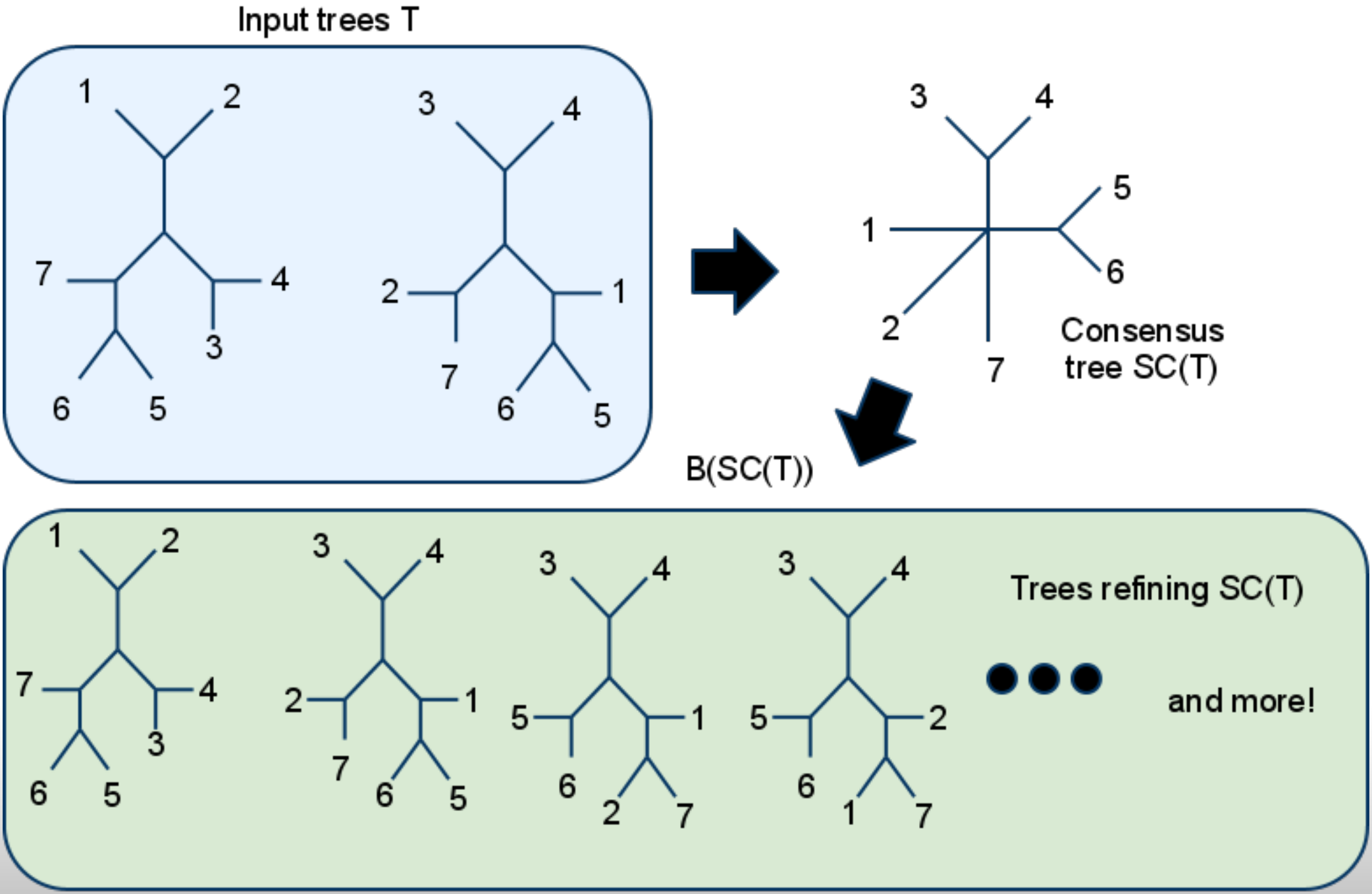


Following containment holds....



Notation:

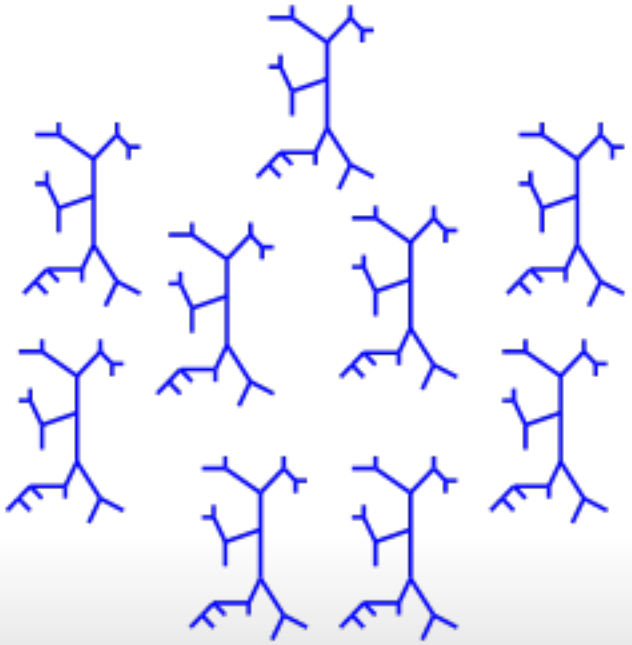
B(t) := {set of all trees refining tree t}



Note: B(SC(T)) contains T, and possibly more trees!

Why compute a single consensus tree?

All trees from phylogenetic analyais



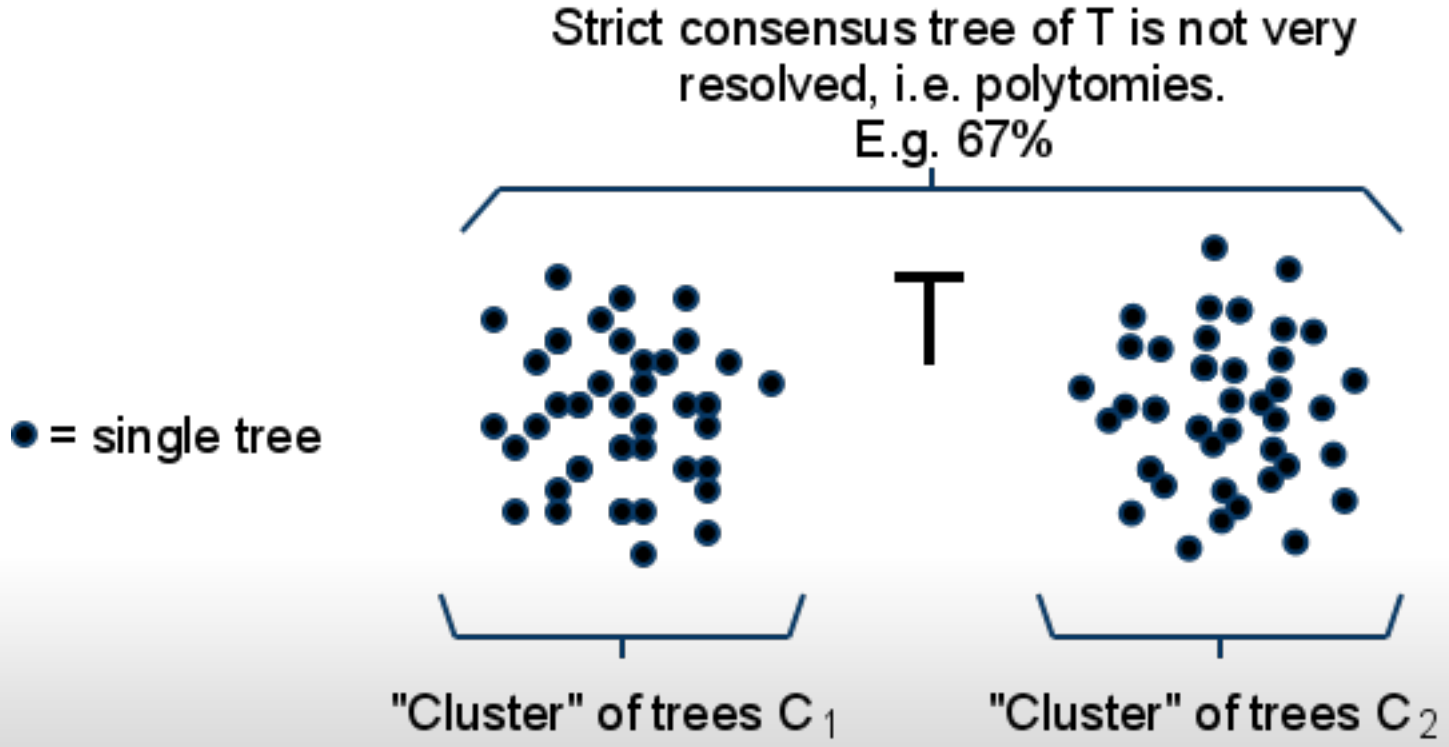
Something in between?

Consensus Tree

Characteristics Trees

Disclaimer: cartoon!

What if our trees **T** from phylogenetic analysis "looked" like:



Characteristic Tree Problem

Fix cluster size k. Let T := { Trees from phylogenetic analysis }.

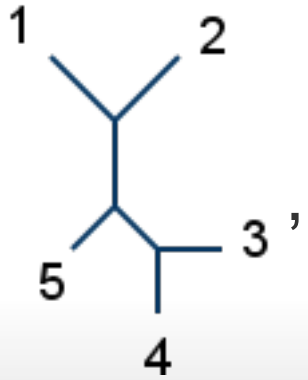
Goal: Cluster trees T into k clusters C_1, C_2, ..., C_k so that...

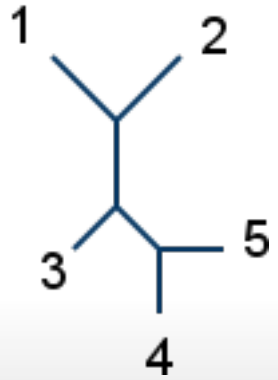
the number of trees refining the strict consensus trees $SC(C_1)$, $SC(C_2)$, ..., $SC(C_k)$ is small.

Need Distance Between Trees

First need a measure of distance between two trees. In Stockham et al. the use the Robinson-Foulds distance.

$d_{RF}(t,t')$ = # edges(splits) in t and not in t' +
edges(splits) in t' and not in t

Ex. $d_{RF}($ 

, 

$) = 1 + 1 = 2$

k-means Clustering

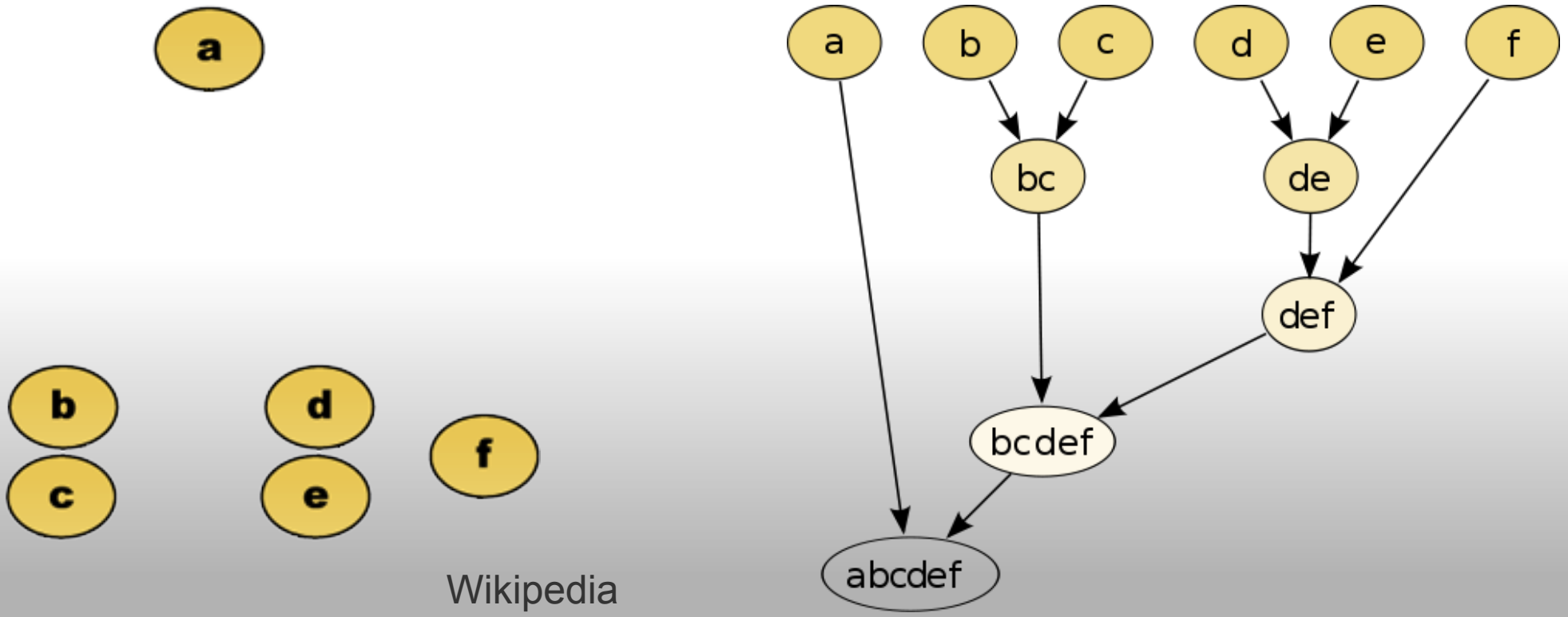
Objective: Cluster data into k clusters such that the trees within each cluster are close as possible to the corresponding cluster mean.

R commercial break!

Agglomerative Clustering

Start with all points in their own cluster. Let k be the desired cluster size. Then proceed...

1. Merge two "closest" clusters.
2. If number of clusters is more than k then goto step 1.



Wikipedia

Stockham et al. did thorough simulation study and found the Agglomerative cluster method best suited for the characteristic tree problem.

Software is available. See manuscript for details.

Data Sets

Caesal: Caesalpine, legume family.

Max parsimony on trnL-trnF intron and spacer regions of chloroplast.
450 trees, 51 leaves. Strict consensus tree SC(**T**) is 77% resolved.

Data Sets

PEVCCA: Stands for Porifea (sea sponge) Echinodermata (sea urchins, sea cucumbers), Vertebrata (fish, reptiles, mammals), Cnidaria (jellyfish), Crustacea (crabs, lobsters, shrimp) and Annelida (roundworms)

Max parsimony on small subunit ribosomal RNA sequences. 5630 trees on 129 leaves divided into 78 phylogenetic islands.

PEVCCA1: contains 168 most parsimonious trees of PEVCCA (1 island). Strict consensus tree 77% resolved.
PEVCCA2: contains 654 next best trees (5 islands). Strict consensus tree is 72% resolved.

`Caesal`

KL(`Agg1`, 5 clusters) → KL(`Agg1`, 3 clusters)

clu	numtrees	specificity	numref
1	108	89.6%	243
2	324	87.5%	729
3	18	89.6%	945
1clu	450	77.1%	8.037×10^{6}

`PEVCCA1`

KL(`Agg1`, 5 clusters) → KL(`Agg1`, 3 clusters)

clu	numtrees	specificity	numref
1	94	92.1%	5.473×10^{7}
2	36	89.7%	2.846×10^{12}
3	38	92.1%	1.148×10^{6}
1clu	168	77.0%	9.264×10^{21}

`PEVCCA2`

KL(`Agg1`, 5 clusters)=21.972959, KL(1 cluster)=53.405270

clu	numtrees	specificity	numref
1	114	92.6%	1.148×10^{7}
2	235	88.1%	7.795×10^{11}
3	6	93.7%	99225
4	211	87.3%	1.465×10^{12}
5	88	86.5%	2.110×10^{10}
1clu	654	72.2%	1.021×10^{26}

Table 2. Comparison of the clustering approach and the single-consensus approach. We use `Agg1` with 3 clusters for `Caesal` and `PEVCCA1`, and 5 for `PEVCCA2`. The 'numtrees' and the 'numref' fields are the number of trees in the cluster and the refinements of the strict consensus of the cluster, respectively. The '1clu' row in each dataset corresponds to the strict consensus of the whole set of trees