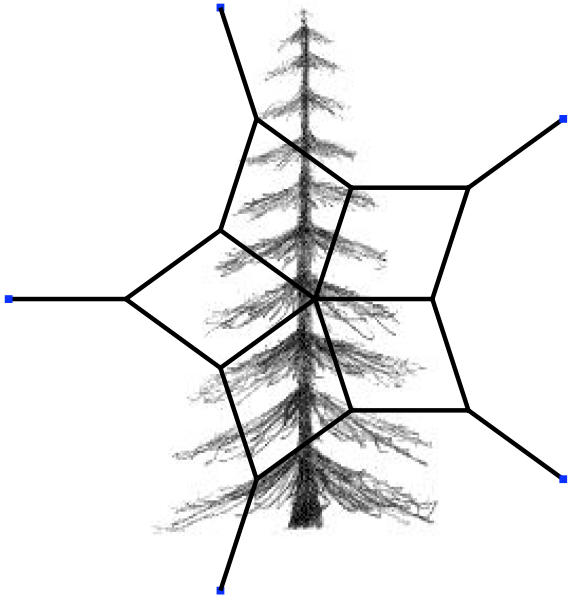
New Software for Phylogenetic Analysis

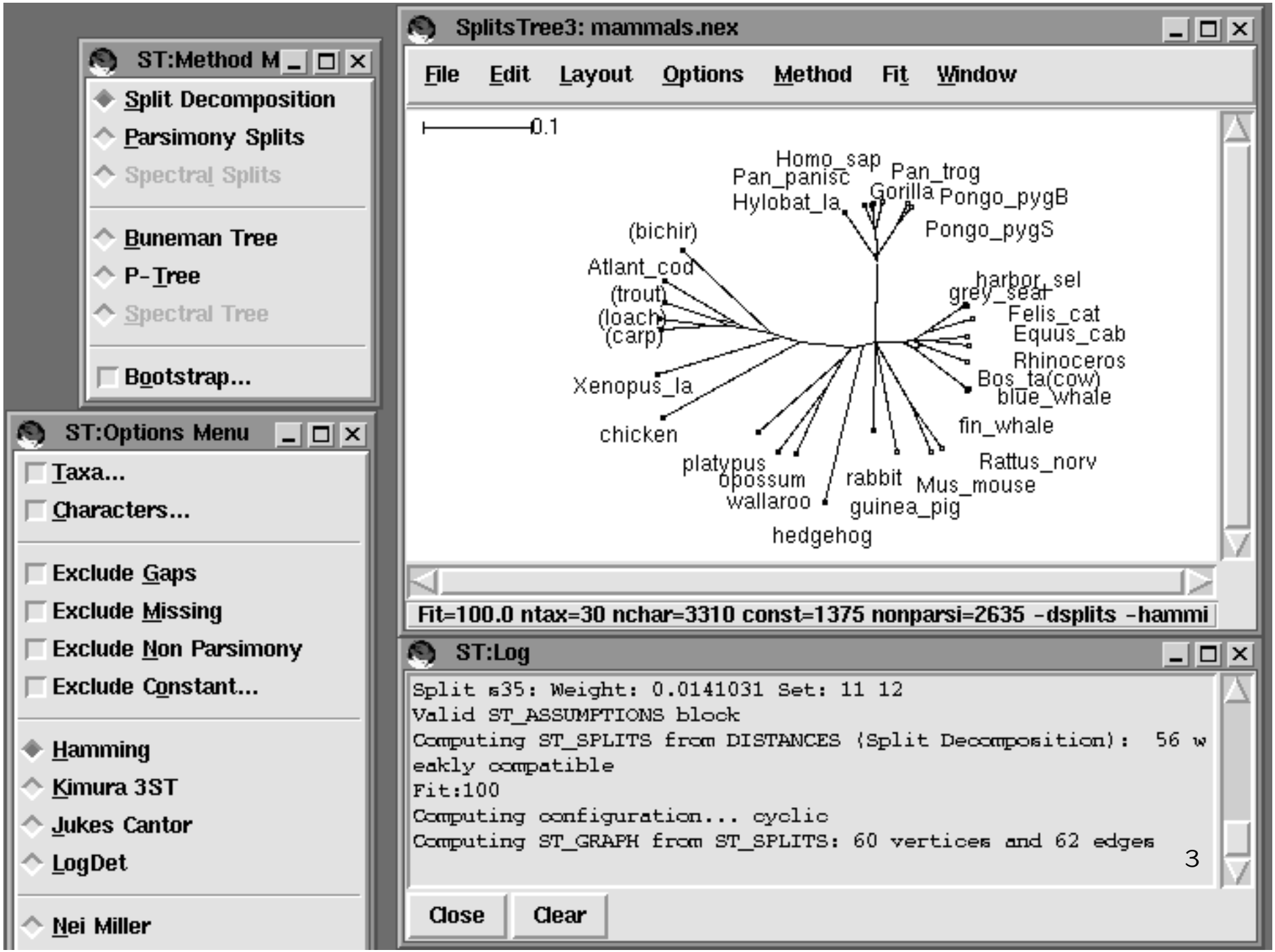


Daniel H. Huson
Applied and Computational Mathematics
Princeton University

huson@math.princeton.edu

Phylogenetic Analysis with SplitsTree

SplitsTree is an interactive program for analyzing and visualizing evolutionary data. Based on the split decomposition method due to H.-J. Bandelt and A.W.M. Dress, it takes as input a distance matrix or a set of aligned sequences and produces as output a graph that represents the evolutionary relationships between the taxa. For ideal data, this graph is a tree, whereas less ideal data gives rise to a more or less tree-like network that can be interpreted as possible evidence for different and conflicting phylogenies. The program supports a number of distances transformations, the computation of parsimony splits, spectral analysis and bootstrapping.


ST:Method M
Split Decomposition
Parsimony Splits
Spectral Splits
Buneman Tree
P-Tree
Spectral Tree
Bootstrap...
ST:Options Menu
Taxa...
Characters...
Exclude Gaps
Exclude Missing
Exclude Non Parsimony
Exclude Constant...
Hamming
Kimura 3ST
Jukes Cantor
LogDet
Nei Miller
SplitsTree3: mammals.nex
File
Edit
Layout
Options
Method
Fit
Window
0.1
Homo_sap
Pan_panisc
Pan_trog
Gorilla
Hylobat_la
Pongo_pygB
Pongo_pygS
(bichir)
Atlant_cod
(trout)
(loach)
(carp)
harbor_sel
grey_seal
Felis_cat
Equus_cab
Rhinoceros
Bos_ta(cow)
blue_whale
fin_whale
Rattus_norv
Mus_mouse
rabbit
guinea_pig
Xenopus_la
chicken
platypus
opossum
wallaroo
hedgehog
Fit=100.0 ntax=30 nchar=3310 const=1375 nonparsi=2635 -dsplits -hammi
ST:Log
Split s35: Weight: 0.0141031 Set: 11 12
Valid ST_ASSUMPTIONS block
Computing ST_SPLITS from DISTANCES (Split Decomposition): 56 w
eakly compatible
Fit:100
Computing configuration... cyclic
Computing ST_GRAPH from ST_SPLITS: 60 vertices and 62 edges
Close
Clear

Trees *are* Compatible Split Systems

Evolutionary relationships are usually represented by a phylogenetic tree T, i.e. a tree whose leaves are labeled by a set X of taxa and whose remaining vertices are unlabeled and of degree at least 3. Any edge e of T defines a split $S = \{A, A'\}$ of X, i.e. a partitioning of X into two non-empty sets A and A', consisting of all taxa on the one side, or the other, of e in T. A system Σ of such splits is called *compatible*, if for any two splits $S_1 = \{A_1, A_1'\}$ and $S_2 = \{A_2, A_2'\}$ in Σ one of the four intersections

$$A_1 \cap A_2, A_1 \cap A_2', A_1' \cap A_2 \text{ or } A_1' \cap A_2'$$

is empty.

Any phylogenetic tree T gives rise to a compatible split system Σ . Vice versa, any compatible split system Σ corresponds to a unique phylogenetic tree T (Buneman 1971). So, tree reconstruction for a given set of taxa X is equivalent to computing a compatible system of splits Σ for X and determining a weight for each split Σ that corresponds to the length of the associated edge.

Weak Compatibility *allows* Networks

To obtain graphs more general than trees, one must consider less restricted systems of splits. Let X be a set of taxa. A split system Σ of X is called *weakly compatible*, if for any *three* splits S_1, S_2, S_3 and all $A_i \in S_i$ $(i = 1, 2, 3)$, at least one of the four intersections

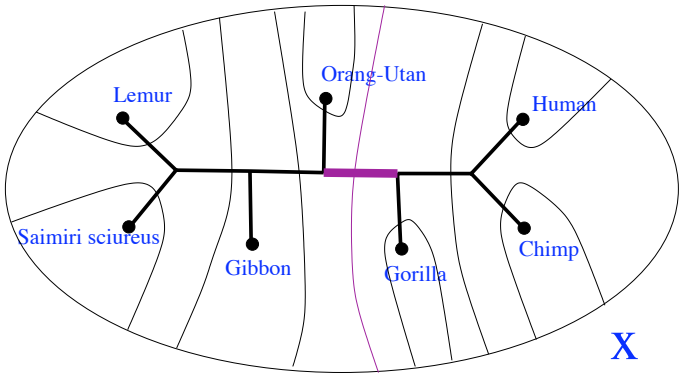
$$A_1 \cup A_2 \cup A_3, A_1 \cup A_2' \cup A_3', A_1' \cup A_2 \cup A_3' \text{ or} A_1' \cup A_2' \cup A_3$$

is empty.

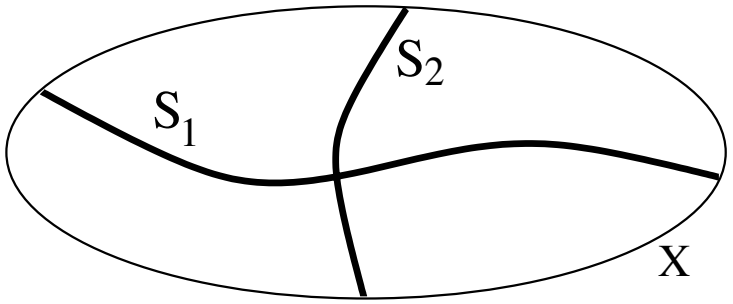
The split decomposition technique defines a unique decomposition of any given distance matrix into a weighted sum of weakly compatible splits that can be efficiently computed (Bandelt and Dress 1992).

A *split graph representing* a weakly compatible split system Σ is a graph $G(\Sigma) = (V, E)$ whose vertices $v \in V$ are labeled by the set of taxa X and whose edges $e \in E$ are straight line segments that represent the splits in Σ.

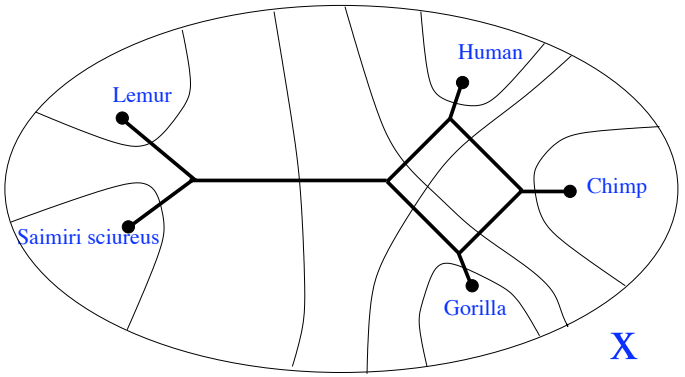
Compatible Splits



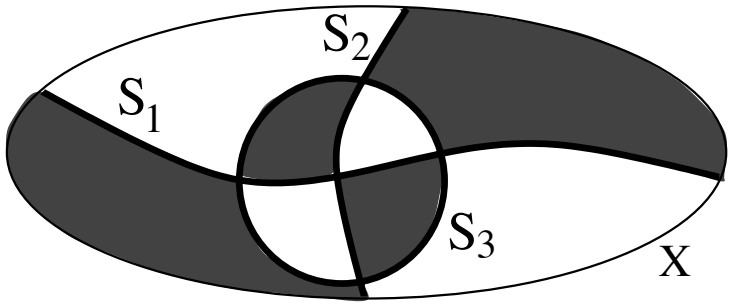
Two splits S_1 and S_2 of X are *compatible*, if one of the four possible intersections is empty:



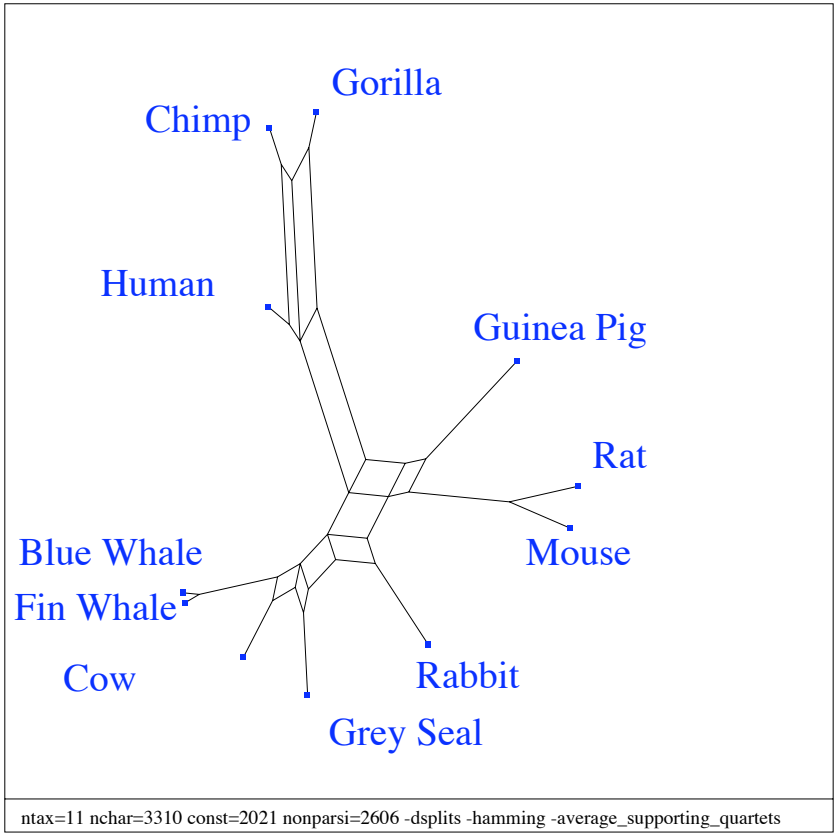
Weakly Compatible Splits



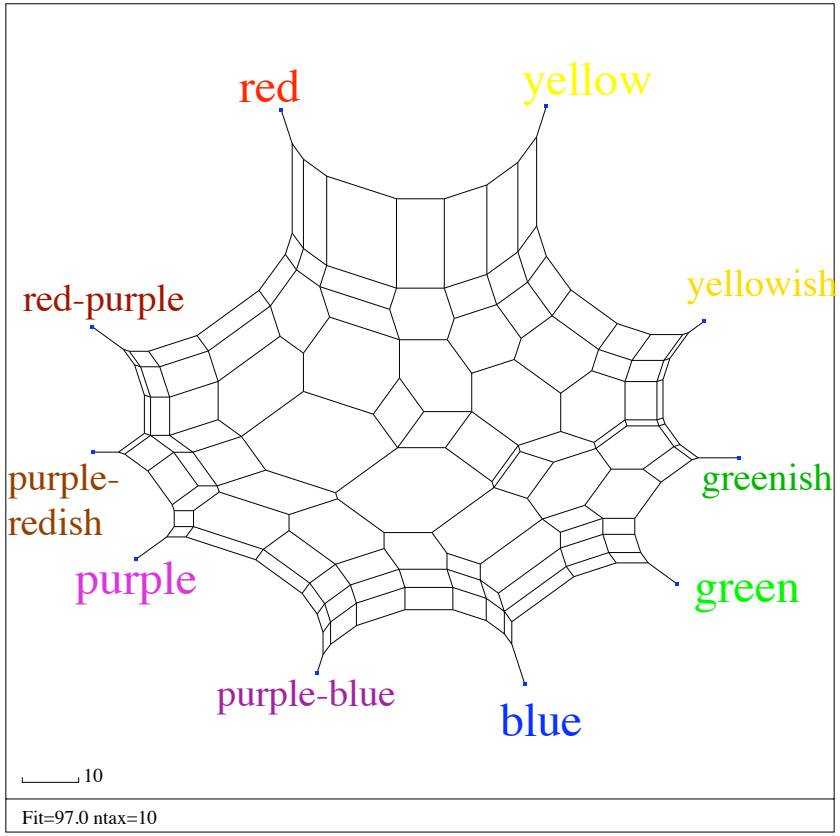
Three splits S_1, S_2 and S_3 of X are *weakly compatible*, if one of the "black" and one of the "white" intersections is empty:



Examples of Split Graphs



Mammalian mtDNA



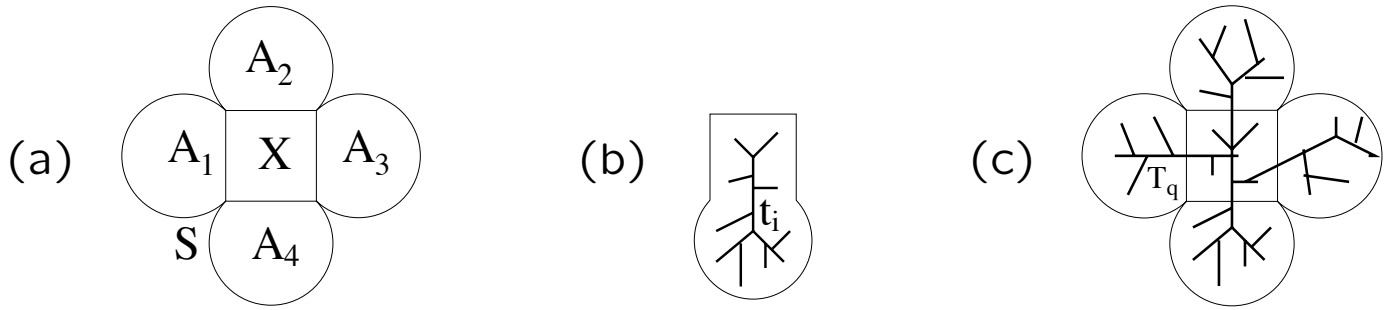
Subjective Color Distances

(Helm 1964)

The Disk-Covering Method

A main drawback of split decomposition is that it only recovers those edges of a tree or graph that are supported by *every* quartet of taxa in the given data set. It suffers from lack of resolution on large and/or very divergent datasets, and will typically only work well on data sets of size up to 20-30, say.

One possible solution to this problem is offered by the *disk-covering method (DCM)* suggested by Tandy Warnow (Huson, Nettles and Warnow 1999). The idea is to compute a phylogenetic tree or network on a large dataset by (1) decomposing the problem into a number of overlapping subproblems that are significantly smaller and less divergent than the original problem, (2) solving each of these subproblems using any base method for building trees or split graphs and then (3) merging the subsolutions to obtain a solution for the whole data set.



This method comes in a number of flavors and can be used in conjunction with any phylogenetic base method such as neighbor-joining, parsimony or split decomposition.

Implementation, Availability and References

SplitsTree is written in C++. The graphical user interface of SplitsTree version 3 is based on Tcl-Tk.

- SplitsTree v2.4.1 for MacOS is available from:
  `ftp://ftp.uni-bielefeld.de/pub/math/splits/splitstree2`

- SplitsTree v3.1 for Windows95 and for Linux is available from:
  `ftp://ftp.uni-bielefeld.de/pub/math/splits/splitstree3`

References:

- H.-J. Bandelt and A.W.M. Dress. A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92:47–105, 1992.

- A.C. Barbrook, C.J. Howe, N. Blake, and P. Robinson. The phylogeny of the Canterbury Tales. *Nature*, 394:839, 1998.

- D.H. Huson. SplitsTree: A program for analyzing and visualizing evolutionary data. *Bioinformatics*, 14(10):68–73, 1998.

- D.H. Huson, S. Nettles, and T.J. Warnow. Obtaining highly accurate topology estimates of evolutionary trees from very short sequences. RECOMB'99, 1999.