Molecular Phylogenetics

By Zhu Hongwei

Introduction

- Brief introduction to phylogenetics basics
- Phylogenetics tree construction methods
- Programs for phylogenetic tree construction

Phylogenetics Basics

- *Phylogenetics*

  the study of the evolutionary history of living organisms

- Tree branching patterns called *dendrogram*

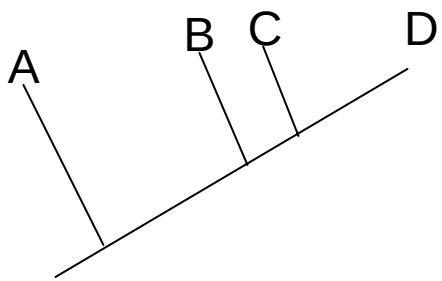
- major assumption for phylogenetics

  1: molecular sequences used in phylogentic construction are homologous

  2: each position in a sequence evolved independently

Phylogenetics Basics

- *Terminology*
- rooted trees & unrooted trees



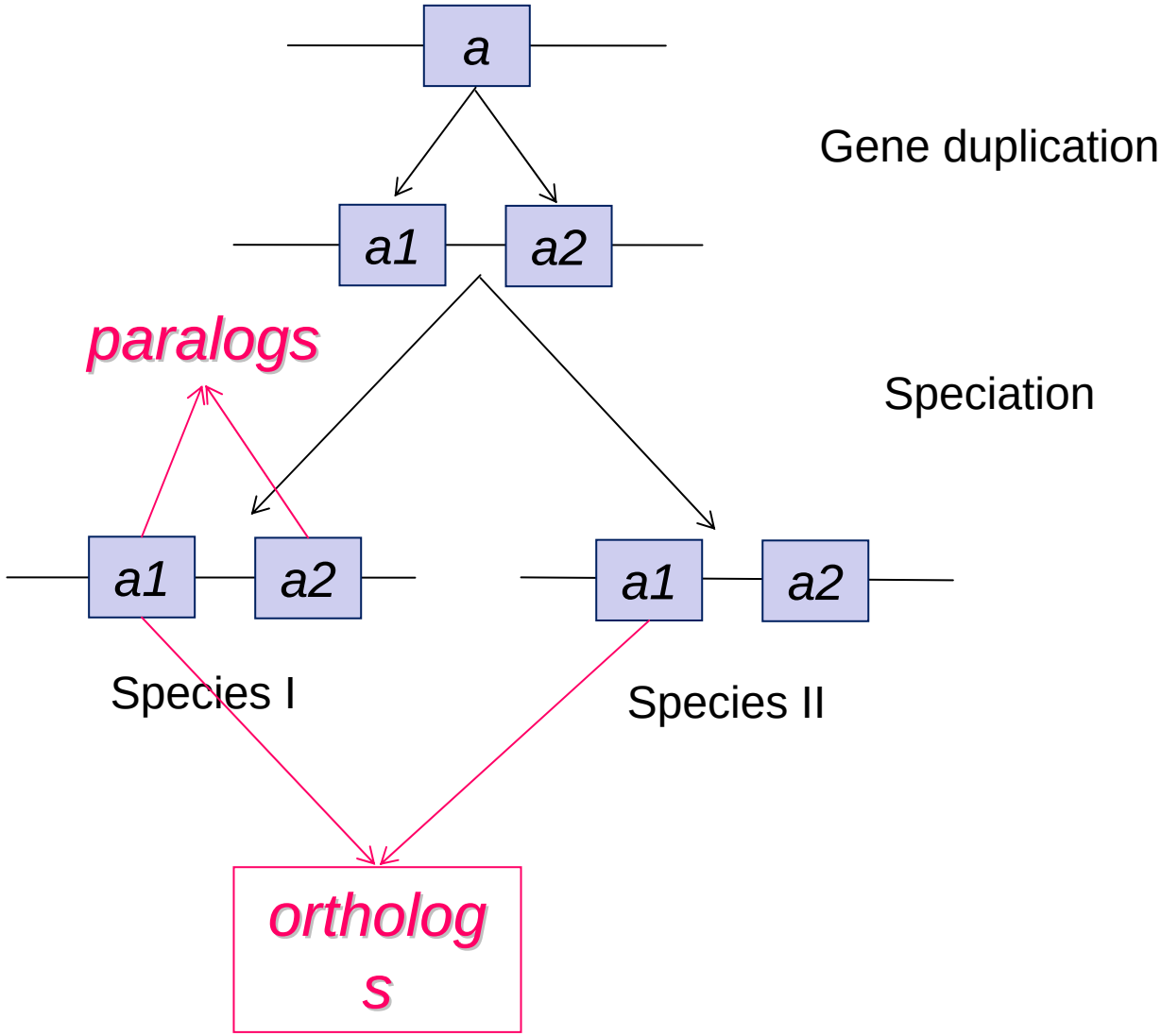
rooted

unrooted

the best way to root a tree is to use an outgroup

- orthologs: Homologous genes evolve from the same ancestral gene
- paralogs: Homologous genes come from gene duplication

Phylogenetics Basics



Phylogenetics Basics

- gene phylogeny versus species phylogeny
- gene phylogeny: phylogeny from a gene or protein sequence

  only describes the evolution of that particular gene/protein

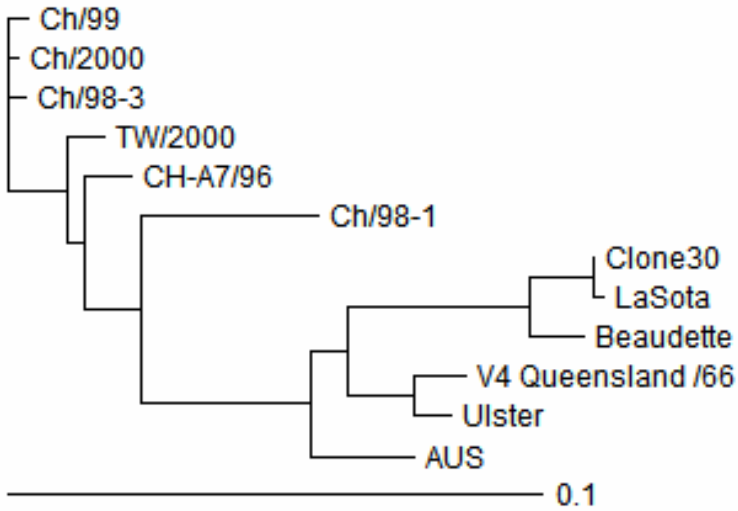
  does not necessarily correlate with the evolutionary path of the species
- In a species tree, the branching point at internal node represents the speciation event

Phylogenetics Basics

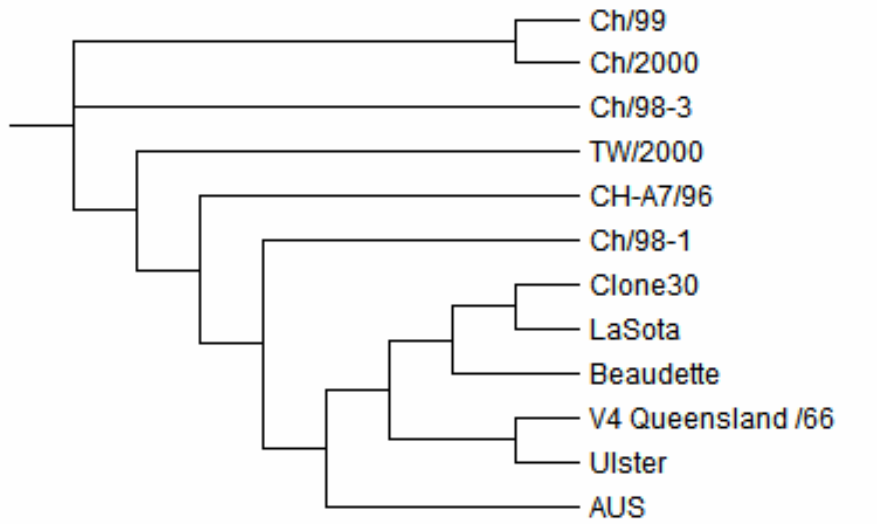
- *Forms of tree representation*
- cladogram trees and phylogram trees
- Phylogram, –scaled, –the branch length represent the amount of evolutionary divergence

  advantages: the trees shows both the evolutionary relationships and information about the relative divergence time of the branches
- Cladogram, –unscaled, –the branches lengths, neatly, but have no phylogenetic meaning
- *Newick format:* (((B,C),A),(D,E)), (((B:1,C:2),A:2),(D:1.2,E:2.5)),

Phylogenetics Basics



Phylogram trees



Cladogram trees

Phylogenetics Basics

- **procedure**
- i: choice of molecular markers

  either nucleotide or protein sequence data can be chosen

  depend on the properties of the sequences and the purposes of the study

  for studying very closely related organisms, nt seq can be used;

  for more widely divergent groups of organisms slowly evolving nt seq(rRNA) or prot seq can be used

Phylogenetics Basics

- **procedure**
- i: choice of molecular markers

  prot seq are preferable in most cases

  reasons:

  prot seq are more conserved as the result of degeneracy

  nt seq are more codon biased than prot seq

  gaps introduced to maximize alignment scores always cause frameshift errors

Phylogenetics Basics

- **procedure**
- ii: alignment

  most critical step in the procedure

  only the correct alignment produces phylogenetic tree

  automatic seq alignment almost always contain errors and should be edited or refined if necessary

  T-Coffee program should be used

  clustalw/clustalx/bioedit/…can also do

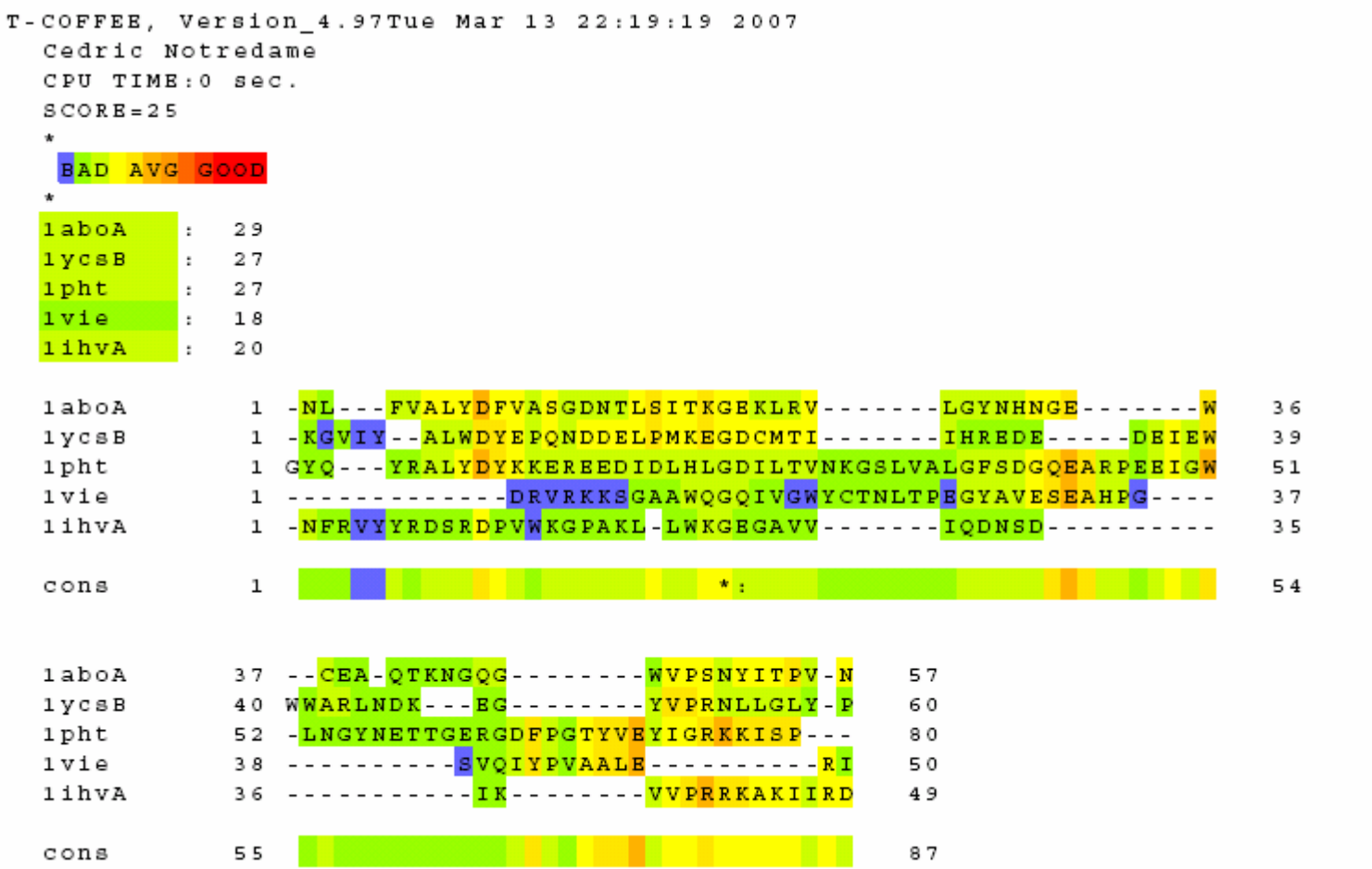
Phylogenetics Basics

- **procedure**
- ii: alignment

  necessary to decide whether to use the full alignment or to extract parts of it, rather subjective

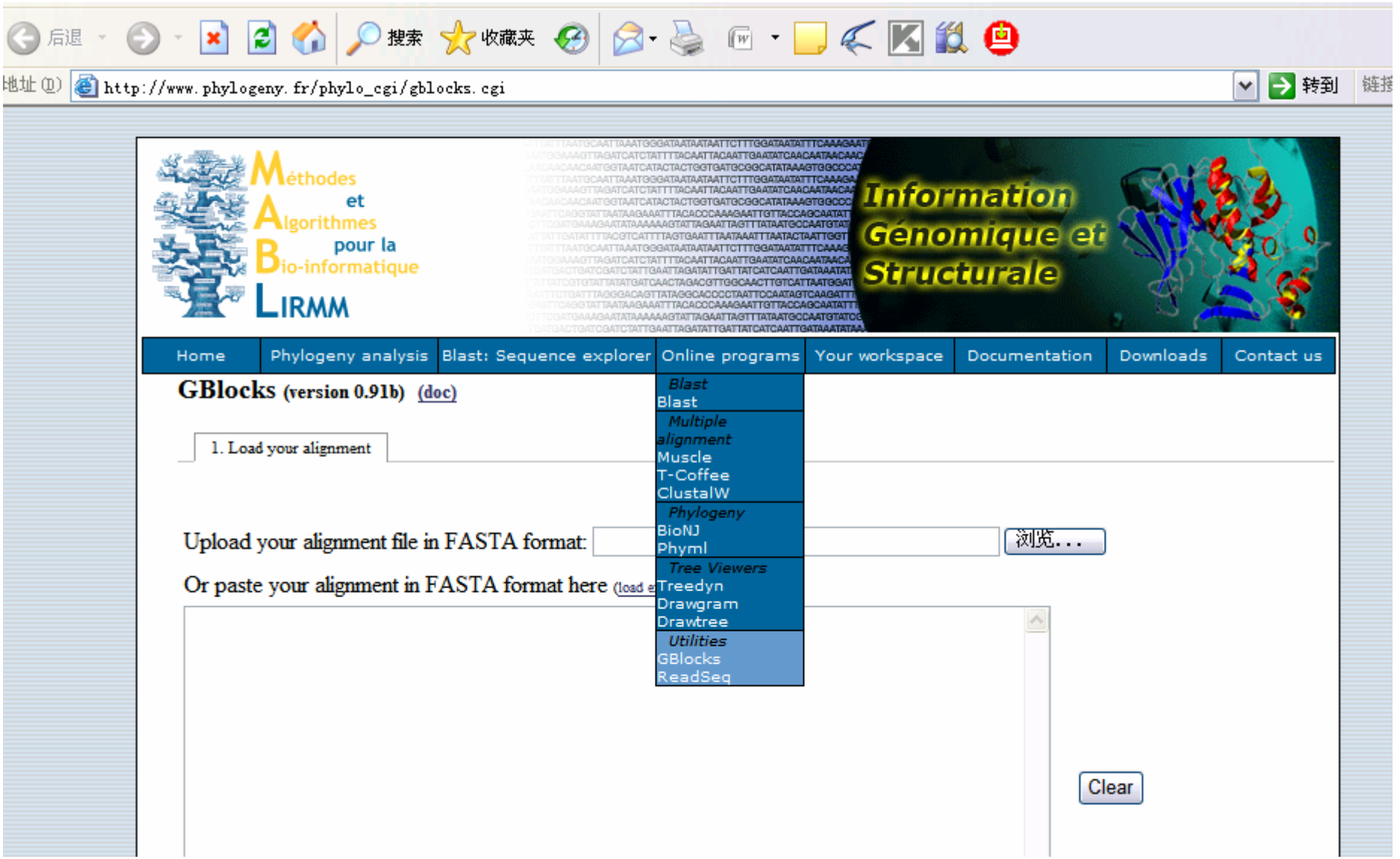
  to improving alignment quality, Rascal and NorMD can help

  the program Gblocks can help to detect and eliminate the poorly aligned positions and divergent regions



back

T-Coffee program demonstration



Gblocks program demonstration

Phylogenetics Basics

- **procedure**
- iii: multiple substitutions

  homoplasy ---multiple substitutions and convergence at individual position obscure the estimation of the true evolutionary distances

  to correct homoplasy, statistical models are needed:

  the commonly used nt substitutions models are Jukes-Cantor and Kimura models

  the commonly used AA substitutions models are PAM and JTT models

Phylogenetics tree construction methods

- The principle of minimum evolution or maximum parsimony often applied
- The two main catalogues: phenetic methods and cladistic methods
- The phenetic methods are distance-based method that measure the pair-wise differences among sequences and build the tree totally from resultant distance matrix
- The cladistic methods are character-based methods, all possible topologies are evaluated and one that is chosen is this that optimizes the evolution

Phylogenetics tree construction methods

- **1: Unweighted pair group method with arithmetic mean (UPGMA)**
- UPGMA is the simplest method of phylogeny
- It uses clustering approach and uncorrected data to build a tree
- Steps for building a tree

- 1. Construct distance matrix
- 2. Cluster the two shortest distance OTUs into an internal nodes
- 3. Recalculate the distance matrix
- 4. Repeat the process until all OTUs are grouped in a single cluster

Phylogenetics tree construction methods

- **Pros and cons**

- Used to construct phylogenic tree of taxa with the relatively constant rate of evolution
- Simple and fast method
- Do not reflect the evolutionary descent
- Extensively used in literature

Phylogenetics tree construction methods

- **2: Neighbor joining (NJ)**
- the phylogenetic tree is constructed from a star-like tree by grouping OTUs with shortest distance of branch length together
- Steps for building a tree

- 1. Start with distance matrix and star-like tree
- 2. Group the two most similar taxa into a node and calculate the branch length
- 3. Recalculate the distance matrix and branch length and construct a new tree
- 4. Repeat the process until only one terminal is present

Phylogenetics tree construction methods

- **Pros and cons**
- Advantages:
  - Relatively rapid, so it is suitable for analyzing a large dataset
  - Calculate the branch length
- Disadvantages:
  - Construct only one possible tree
  - Yield a biased tree under some condition
  - Compress sequence information

Phylogenetics tree construction methods

- **Pros and cons of phenetic methods**
- Both UPGMA and NJ base on distance matrix to reflect evolutionary relationship
- they compress sequence information into single number, cannot reflect the changes of character states of sequences
- UPGMA and NJ are relatively fast, suitable for analyzing large data set that are not very strong similar
- In general, NJ gives better result than UPGMA

Phylogenetics tree construction methods

- **Cladistic methods**
- Cladistic methods assume that a set of sequences descended from a common ancestor by mutated and selected processes without hybridization or other horizontal gene transfers
- **1: Maximum parsimony(MP)**

- Maximum parsimony assumes that trees with the minimum number of evolutionary changes are the most preferable trees
- MP bases on the number of character-state changes to construct all possible trees and give each a score

Phylogenetics tree construction methods

- Steps for building a tree
- 1. Start with multiple alignment
- 2. Construct all possible topologies and base on evolutionary changes to score each of these topologies
- 3. Choose a tree with the fewest evolutionary changes as the final tree
- **Pros and cons**
- Advantages:
- Reflect the ancestral relationship
- Use all known evolutionary information
- Faster than Maximum likelihood

Phylogenetics tree construction methods

- Disadvantages:
- Yield little information about branch length
- Require long computation time
- Yield biased tree under some conditions

- **2: Maximum likelihood(ML)**

- Maximum likelihood use statistical tool to evaluate a hypothesis about evolutionary history
- It constructs all possible trees of evolutionary history from an observed data set

Phylogenetics tree construction methods

- Steps for building a tree
- 1. Start with a multiple alignment.
- 2. List all possible topologies of each data partition (i.e., column)
- 3. Calculate probability of all possible topologies for each data partition.
- 4. Combine data partitions
- 5. Identify tree with the highest overall probability at all partitions as most likely phylogeny
- **Pros and cons**
- Advantages:
- More accurate than other methods. It is often used to test an existing tree.

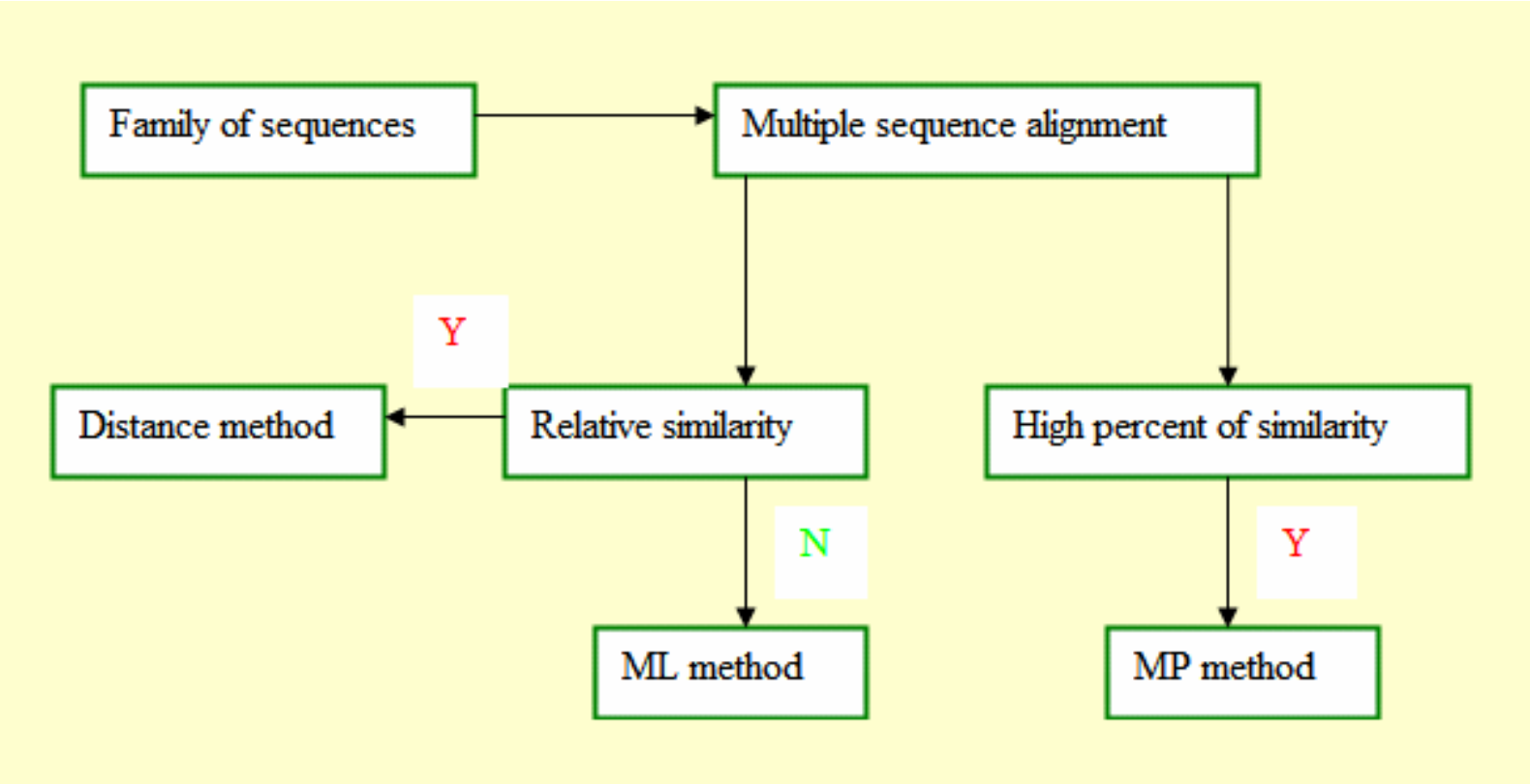
Phylogenetics tree construction methods

- Advantages:
- All the sequence information is used
- Evaluate all possible trees
- Sampling errors have least effect on the method
- Disadvantages:
- Extremely slow
- Impractical for analyzing large data set

Phylogenetics tree construction methods

- **Evaluation of different methods**
- Up to date, None of the tree-building methods make sure to reflect correctly the evolutionary relationship of a sequence set
- There is a recommended phylogeny flowchart to choose right methods

Phylogenetics tree construction methods



Recommended Phylogeny Flowchart

Phylogenetics tree construction methods

- **Phylogenetic tree evalution**
- Boostrapping & Jackknifing
- Bootstrapping: repeatedly sampling trees through slightly perturbed dataset
- Bootstrap results should be interpreted with caution (does not assess the accuracy of a tree, only consistency and stability indicated
- A tree should be bootstrapped 500-1000 times, impossible for MP and ML

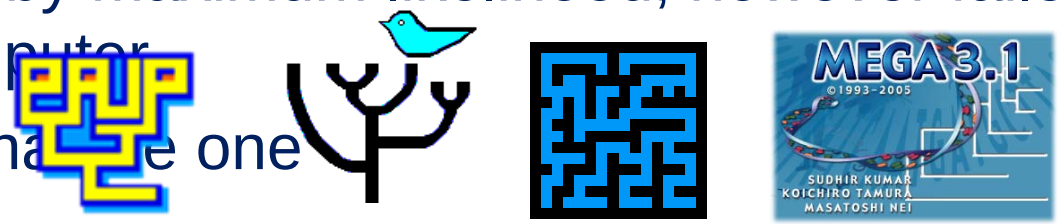
Phylogenetics tree construction methods

- Jackknifing: one half of the sites in a dataset are randomly deleted
- Computing time is much shortened
- The results may not be comparable with that from bootstrapping

Tree Constructing Programs

- There are plenty of programs available for constructing phylogenetic trees
- The most commonly used are as followed:
- PAUP(Phylogenetic analysis using parsimony), one of the most widely used

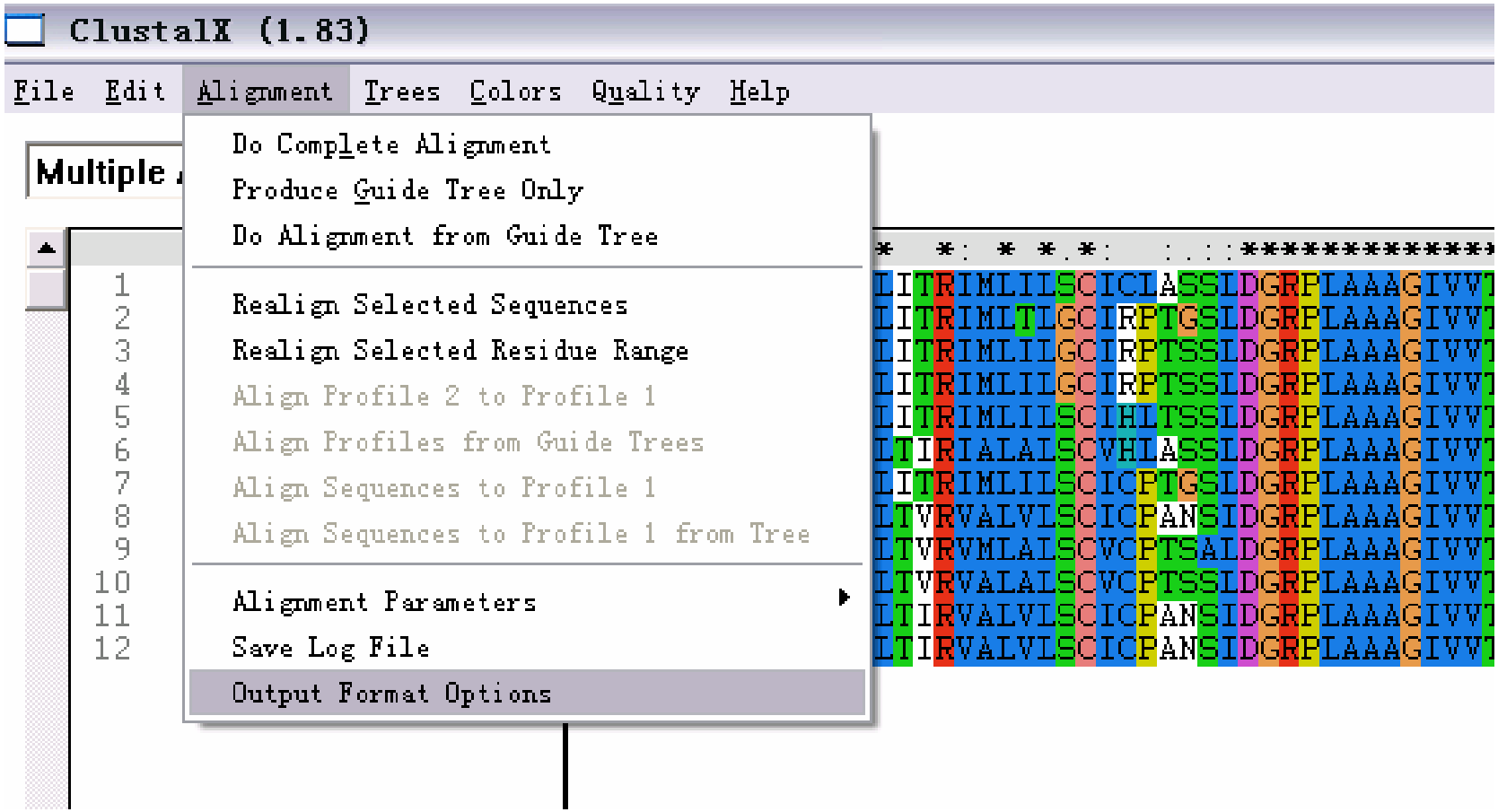
  Commercial phylogenetic package for Macintosh (UNIX version available)
- Phylip package is a free program and can also available online
- TREE PUZZLE is a program that allows various substitution models by maximum likelihood, however failed to install on my computer
- MEGA is a easy to handle one



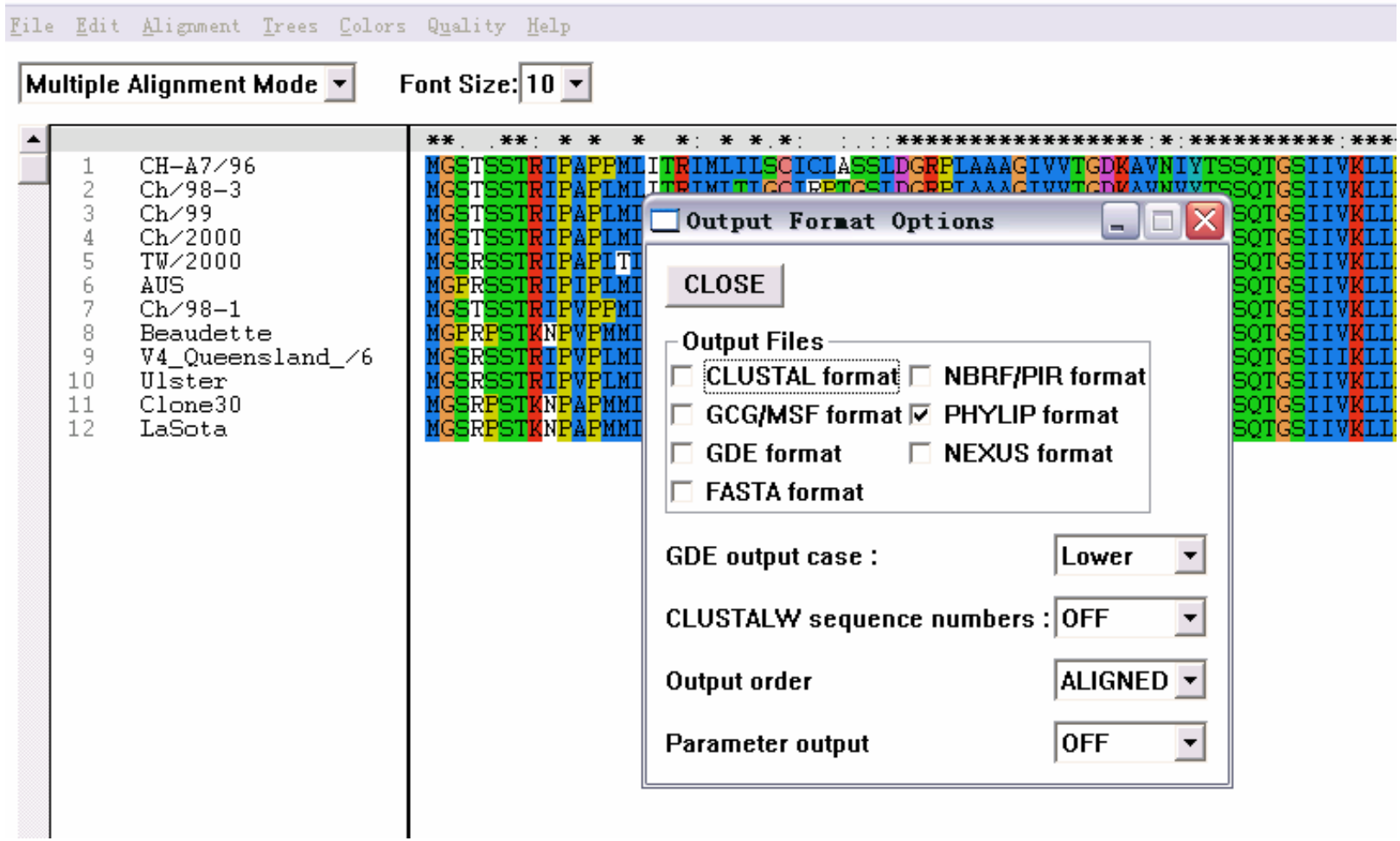
Tree Constructing Programs

- How to use Phylip 
- i: Choose a molecular marker
- ii: Do careful alignment
- iii: Run Phylip program package, select program interest

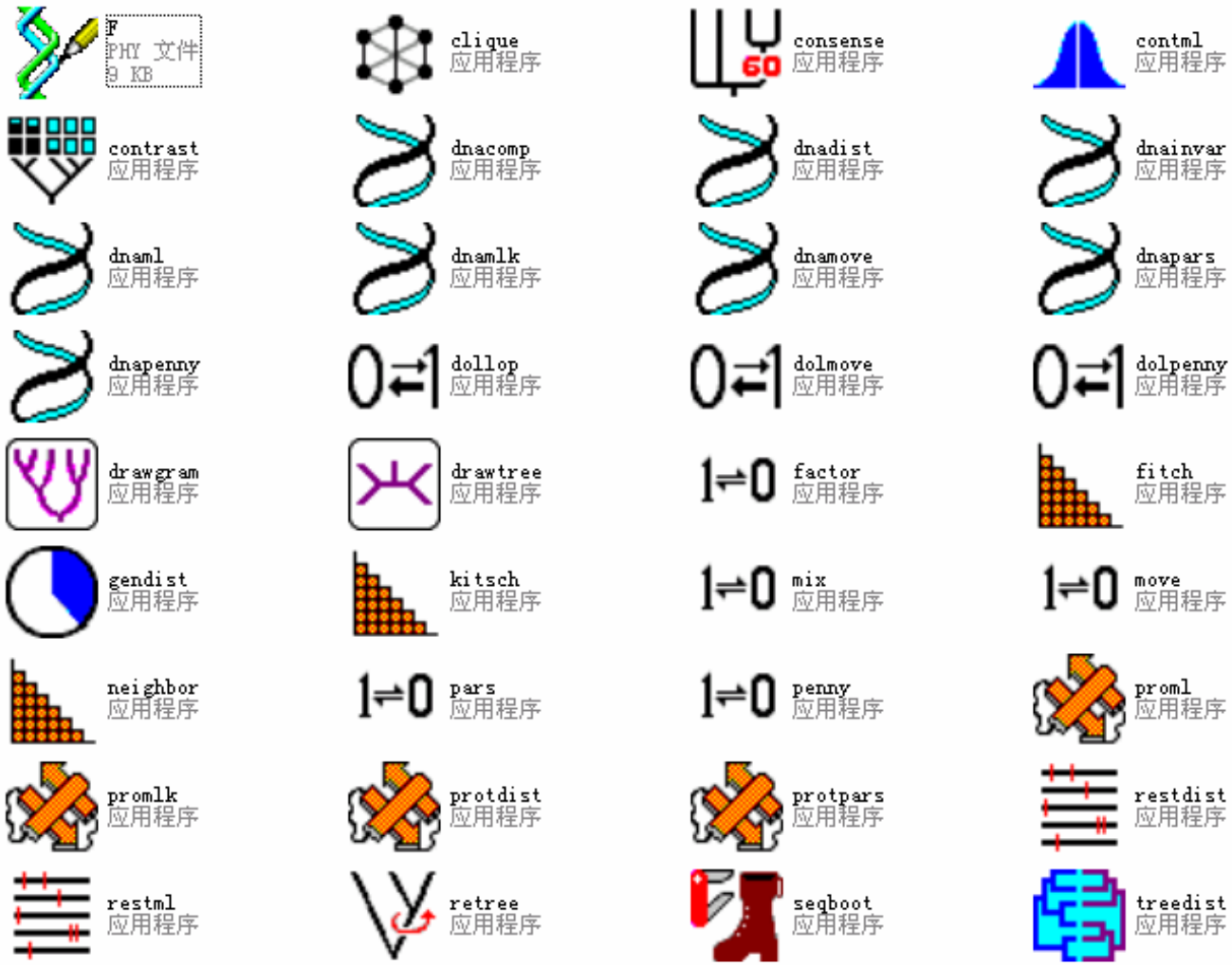
Tree Constructing Programs



Tree Constructing Programs

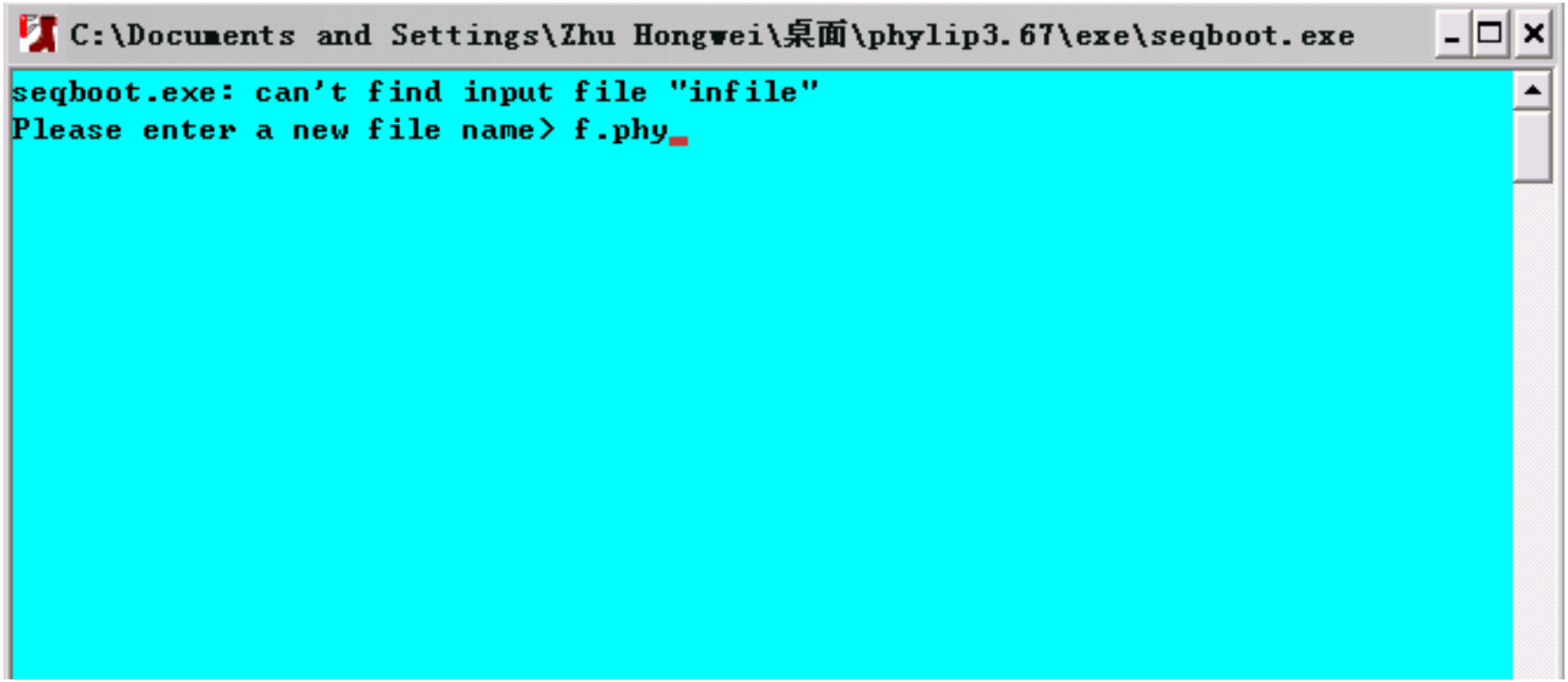


Tree Constructing Programs



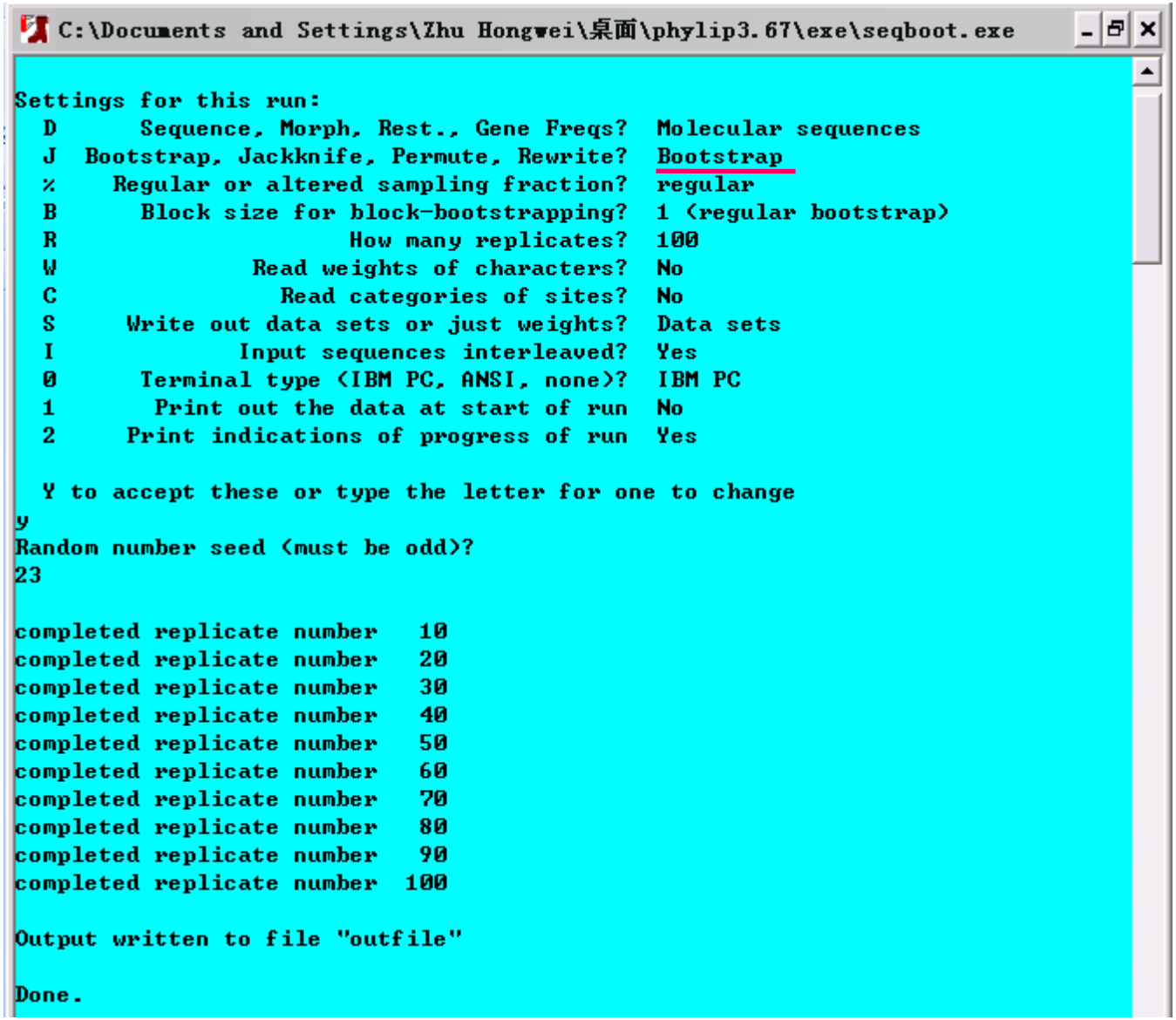
Phylip program package

Tree Constructing Programs



Tree Constructing Programs

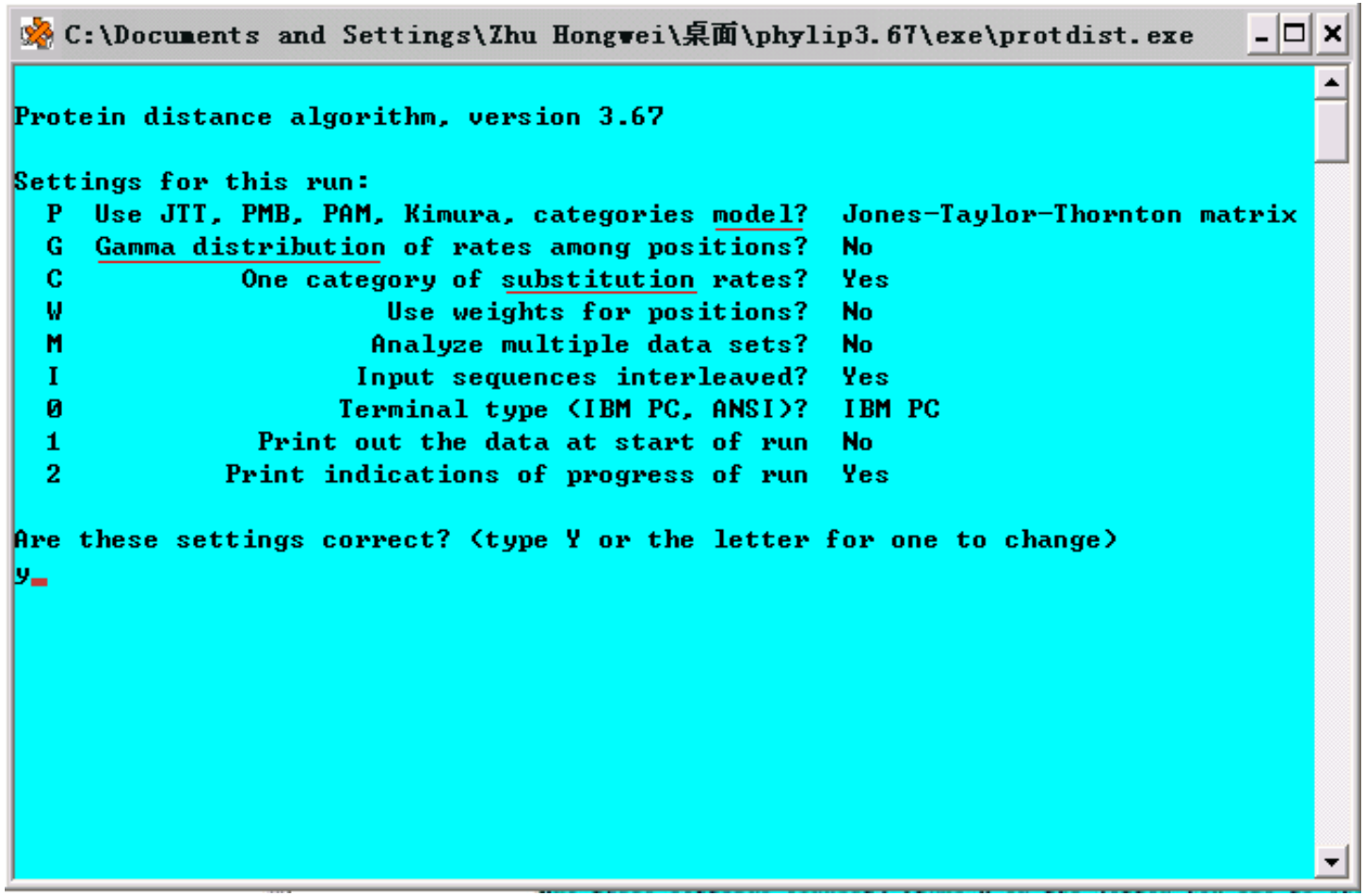
Tree Constructing Programs



Tree Constructing Programs

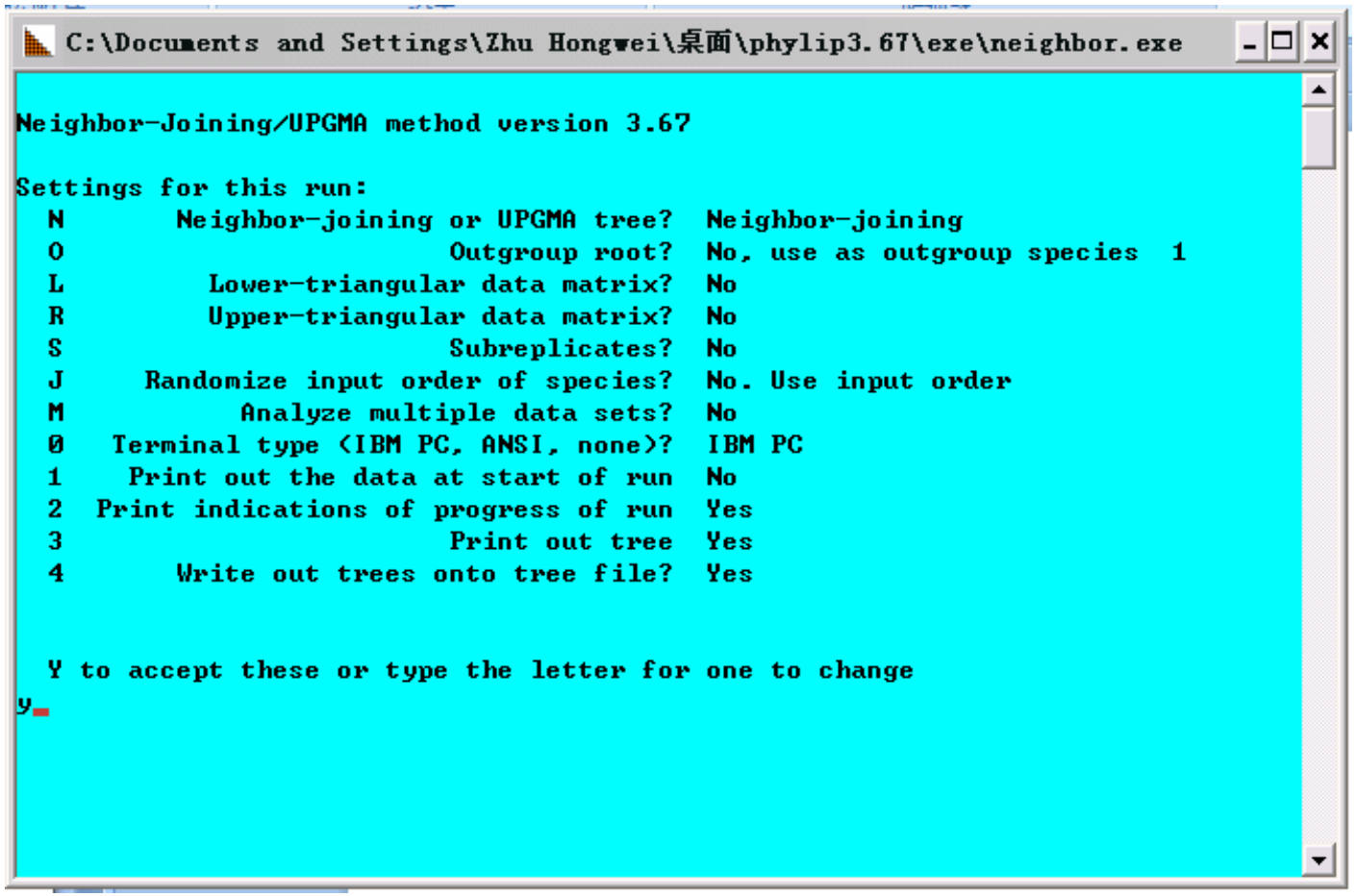
Rename the ‘outfile’ to ‘infile’

Tree Constructing Programs



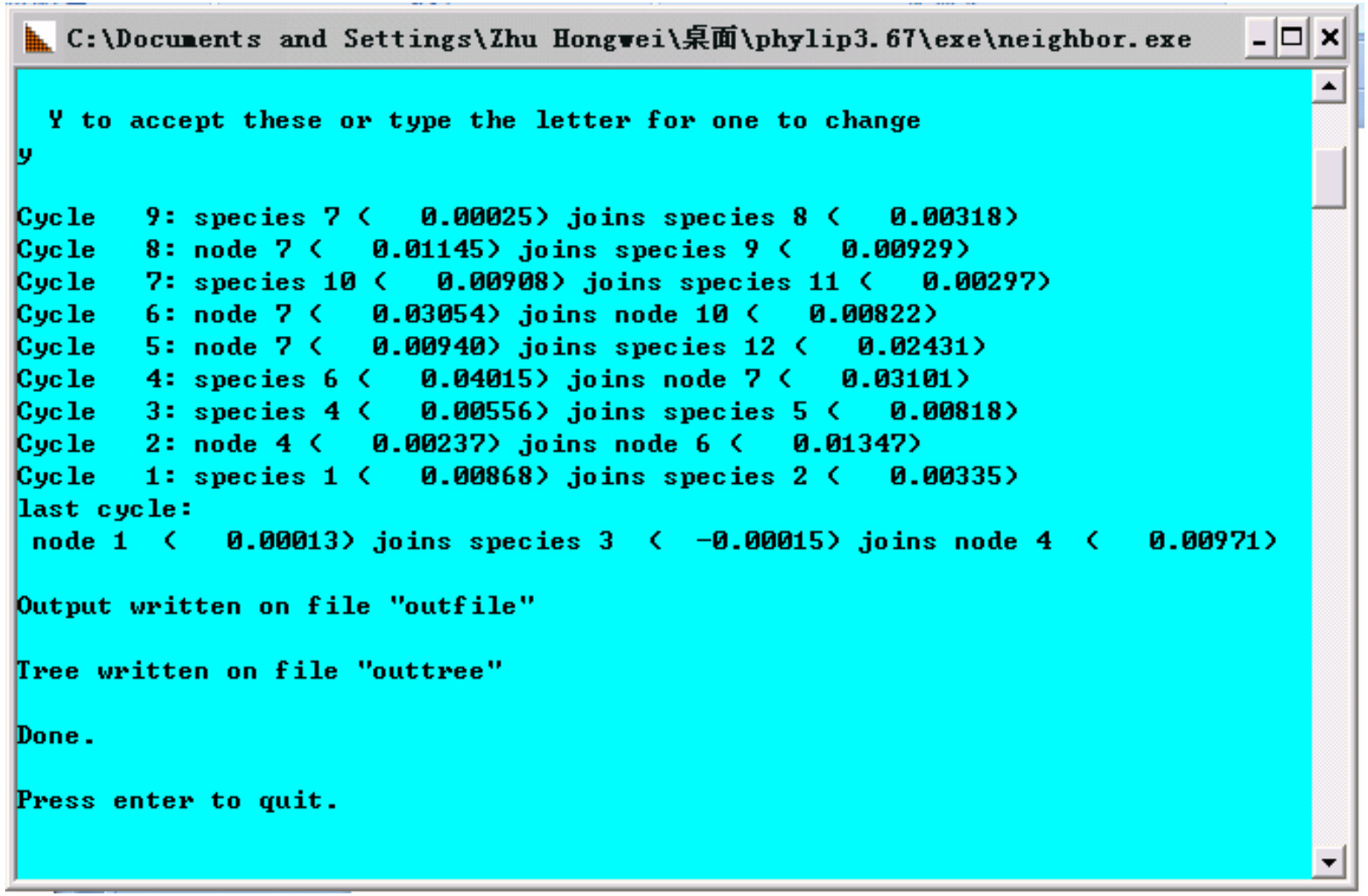
Run ‘protdist.exe’ and an output of ‘outfile’

Tree Constructing Programs

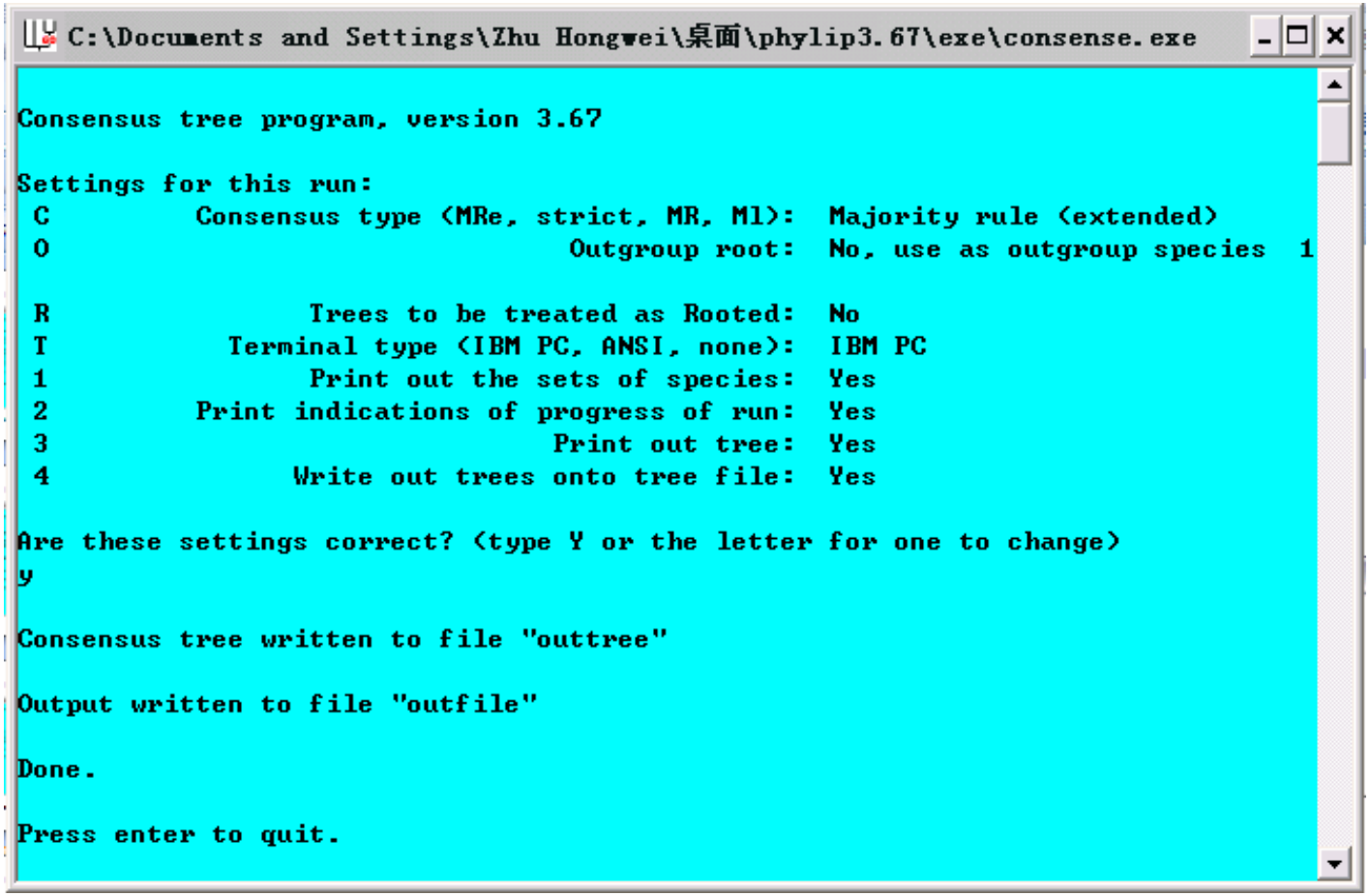


Change ‘outfile’ to ‘infile’ and run ‘neighbor.exe’

Tree Constructing Programs



Tree Constructing Programs



Rename ‘outfile’ to ‘infile’,’outtree’to ‘intree’ and run ‘consense.exe’

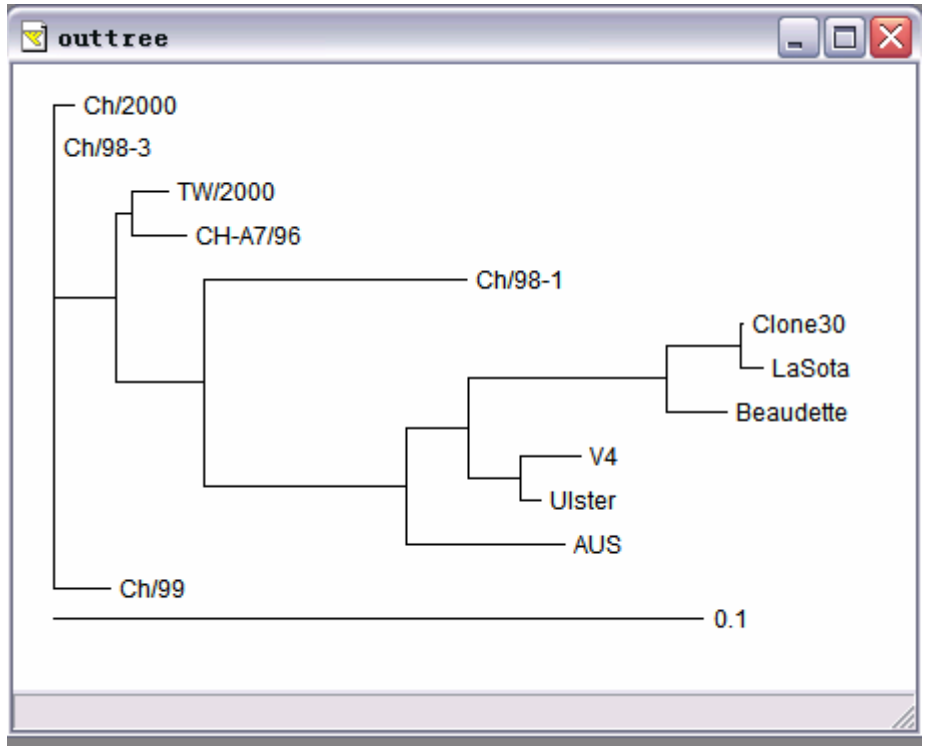
Tree Constructing Programs

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

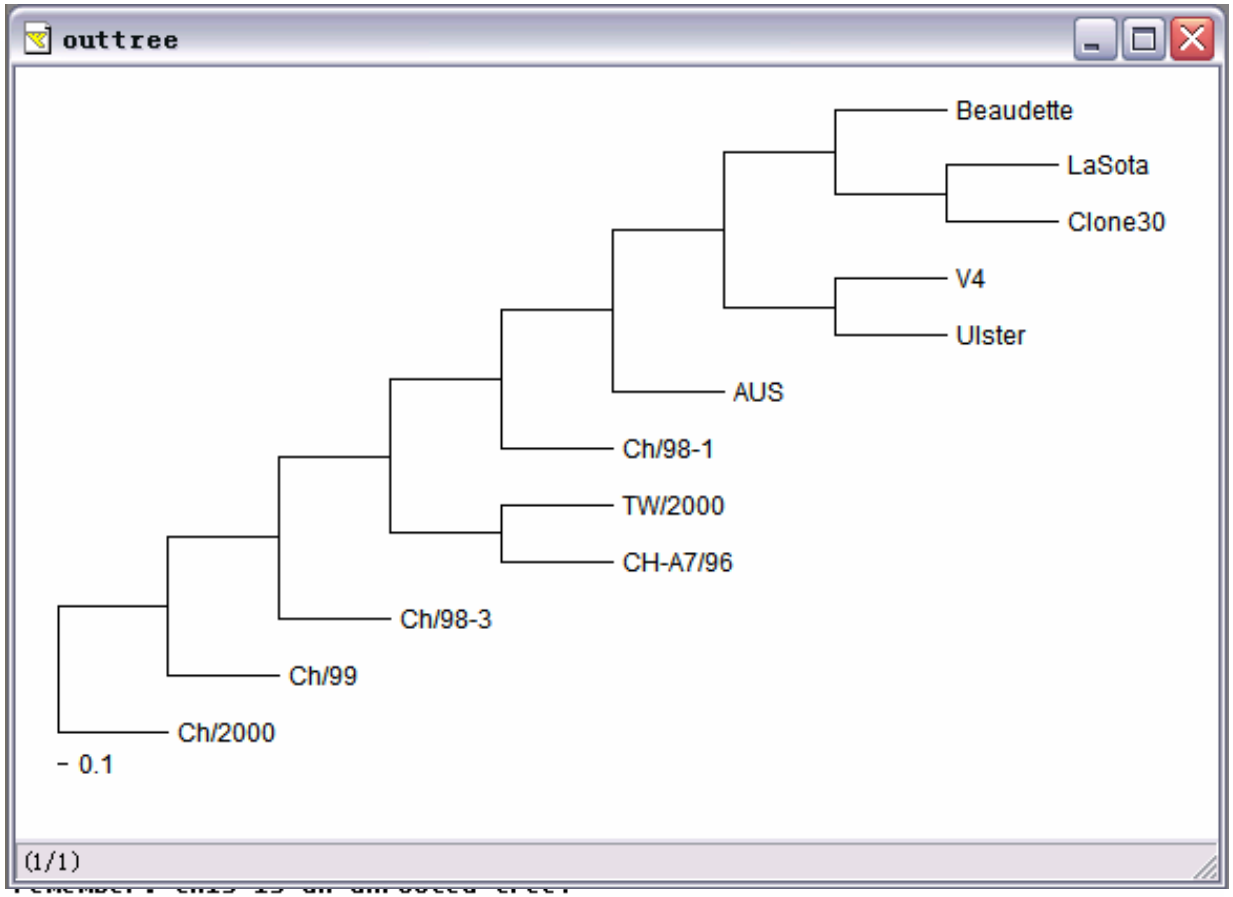
```
Neighbor-joining method

Negative branch lengths allowed


 +Ch/2000
 !
 !  +Ch/98-3
 !  !
 !  !   +TW/2000
 9-10 +-7
 !  ! ! +CH-A7/96
 !  ! !
 !  +-8 +--Ch/98-1
 !     ! !
 !     ! !        +Clone30
 !     ! !      +-1
 !     +-6    +-2 +LaSota
 !       !    ! !
 !       ! +-4 +Beaudette
 !       ! ! !
 !       ! ! ! +V4
 !       +-5 +-3
 !         !   +Ulster
 !         !
 !         +-AUS
 !
 +Ch/99
```



Tree Constructing Programs



There are so many thing covering this!

keep one thing in mind: BE CAUTIOUS when dealing with phylogenetic trees, it is very difficult to find a TRUE tree !

References:

Jin Xiong, Essential bioinformatics, Cambridge University Press, 2006: 127-169

www.genecool.com/bbs/thread-8918-1-1.html

http://www.phylogeny.fr/phylo_cgi/gblocks.cgi

http://www.ebi.ac.uk/t-coffee/

Jean-Michel Claverie, Cedric Notredame, Bioinformatics for dummies (2nd Edition) Wilry Press, 2006

Thanks!!