

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

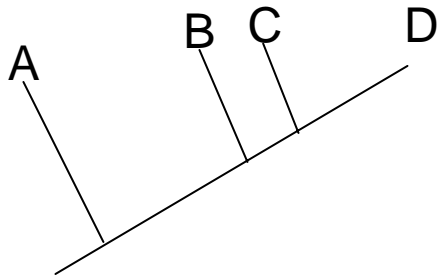
1: molecular sequences used in phylogenetic construction are homologous

2: each position in a sequence evolved independently

Phylogenetics Basics

Terminology

rooted trees & unrooted trees



rooted



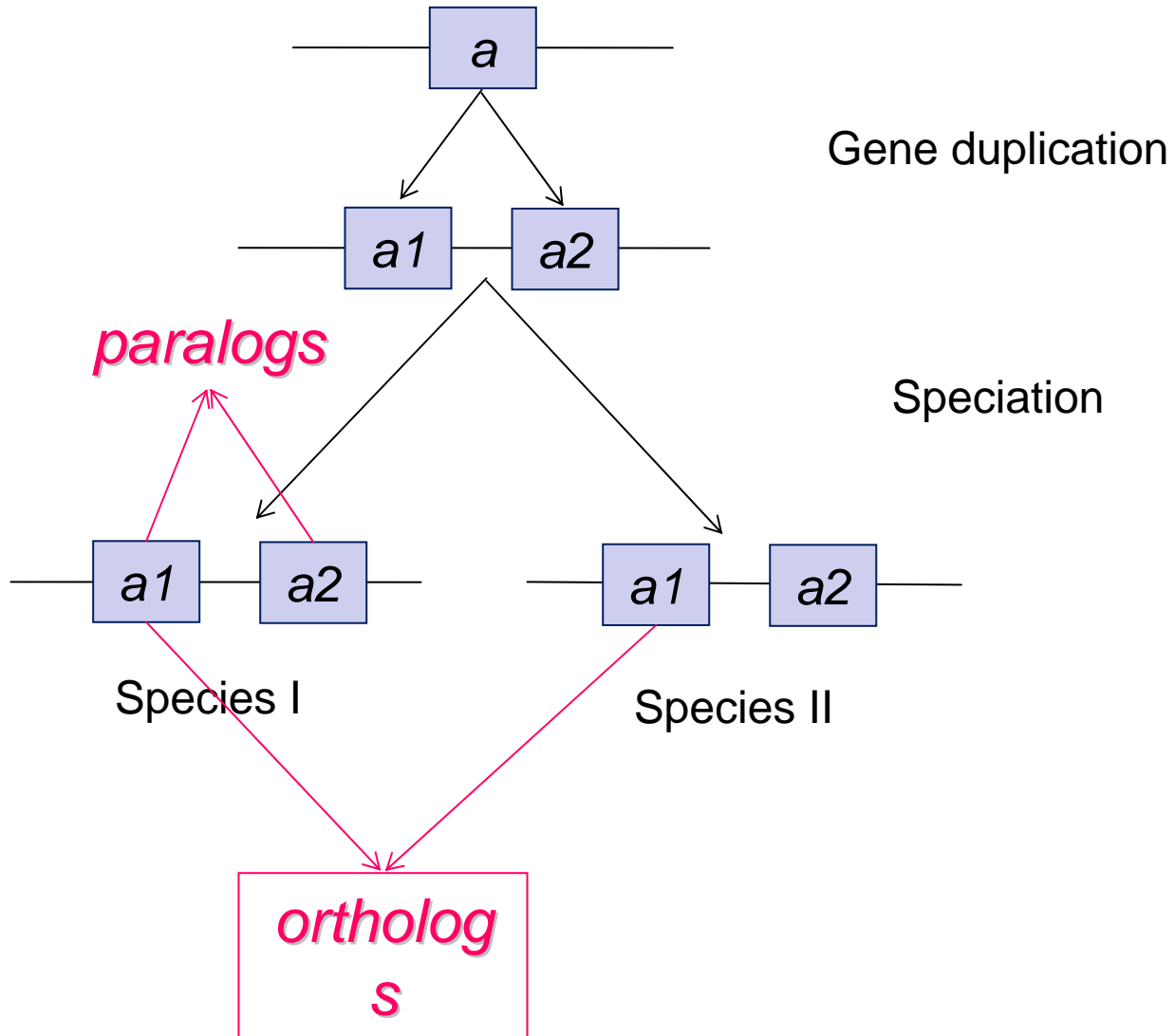
unrooted

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

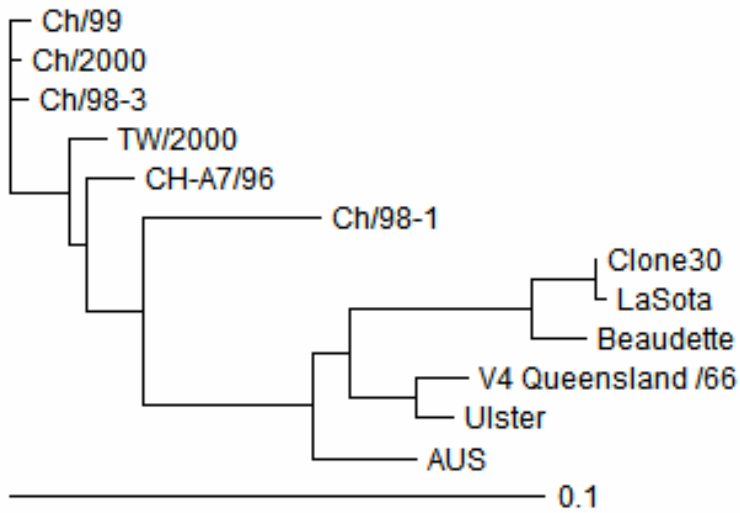
only describes the evolution of that particular gene/protein
does not necessarily correlate with the evolutionary path
of the species

- ✚ In a **species tree**, the branching point at internal node represents the speciation event

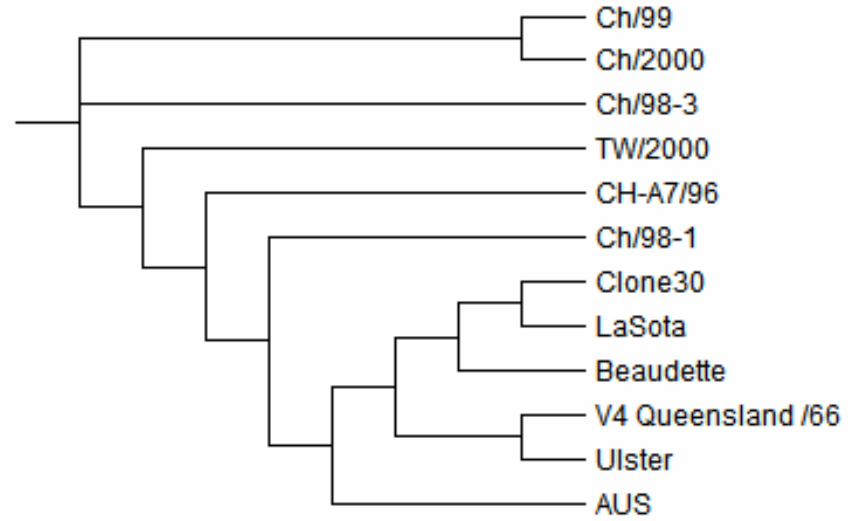
Phylogenetics Basics

- ✚ *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

T-COFFEE, Version_4.97Tue Mar 13 22:19:19 2007

Cedric Notredame

CPU TIME:0 sec.

SCORE=25

*

BAD AVG GOOD

*

laboA : 29
lycsB : 27
lpht : 27
lvie : 18
lihvA : 20

laboA	1	-NL---	FVALYDFVASGDNTLSITKGEKLRV	-----	LGYNHNGE	-----	W	36
lycsB	1	-KGVII	--ALWDYEPQNDDELPMKEGDCMTI	-----	IHREDE	-----	DEIEW	39
lpht	1	GYQ---	YRALYDYKKEREEDIDLHLGDILTVNKGSLVAL	GFSDGQEARPEEIGW				51
lvie	1	-----	DRVRKKS	GAAWQQQIVGWYCTNLTP	EGYAVESEAHFG	----		37
lihvA	1	-NFRVYYRDSRDPVW	KGPAKL	-LWKGBGAVV	-----	IQDNSD	-----	35
cons	1				*			54

laboA	37	--	CEA-QTKNGQG	-----	WVPSNYITPV	-N	57
lycsB	40	WWARLNDK	---EG	-----	YVPRNLLGLY	-P	60
lpht	52	-LNGYNETTGERGDFP	PGTYVE	YIGRKKISP	---		80
lvie	38	-----	SVQIYPVAALE	-----	RI		50
lihvA	36	-----	IK	-----	VVPRRKAKIIRD		49
cons	55						87

[back](#)

T-Coffee program demonstration



Home	Phylogeny analysis	Blast: Sequence explorer	Online programs	Your workspace	Documentation	Downloads	Contact us
------	--------------------	--------------------------	-----------------	----------------	---------------	-----------	------------

GBlocks (version 0.91b) (doc)

1. Load your alignment

Upload your alignment file in FASTA format:

浏览...

Or paste your alignment in FASTA format here (load e

Clear

- Blast
- Blast
- Multiple alignment
- Muscle
- T-Coffee
- ClustalW
- Phylogeny
- BioNJ
- Phyml
- Tree Viewers
- Treedyn
- Drawgram
- Drawtree
- Utilities
- GBlocks
- ReadSeq

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 phylogenetic 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

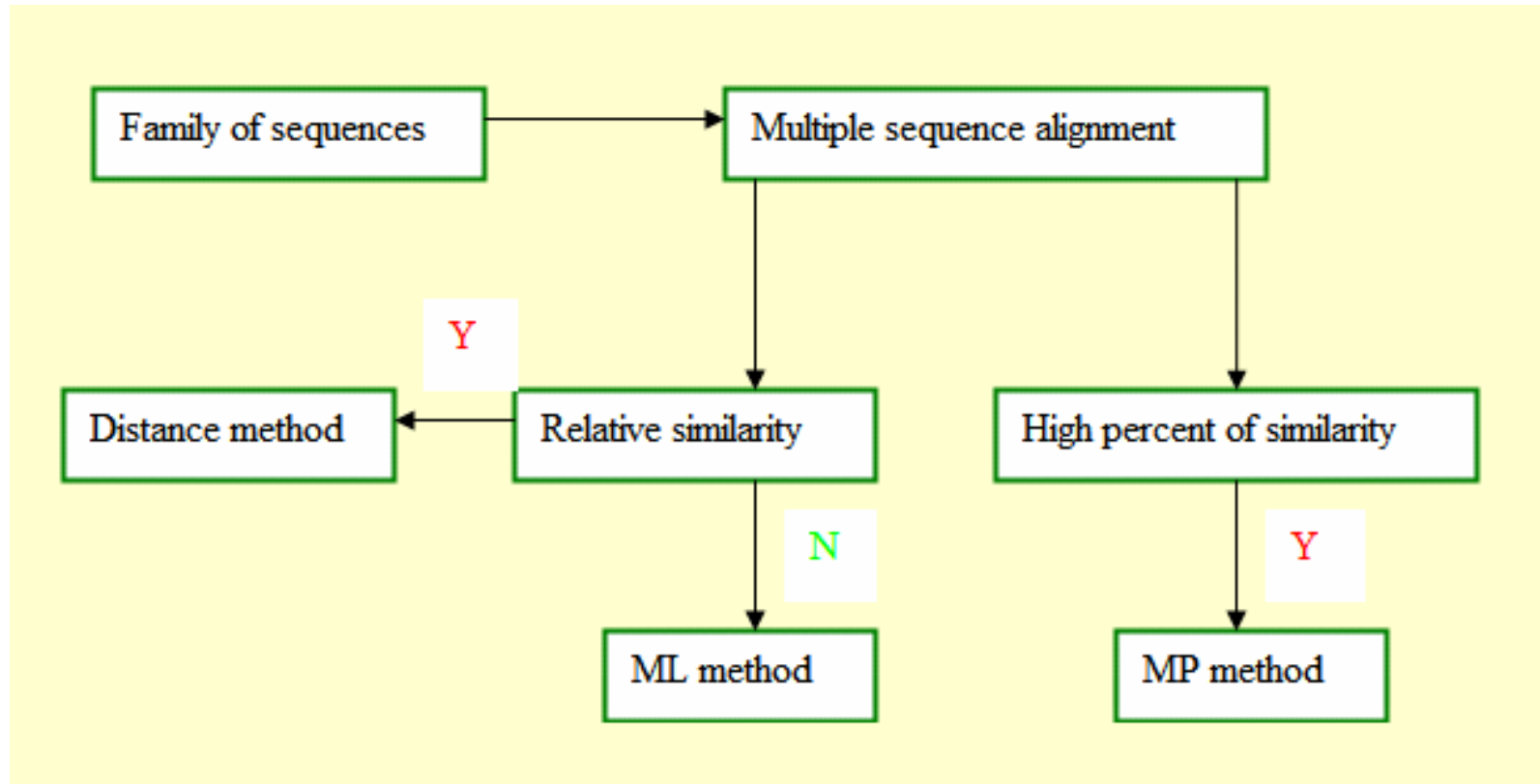
- Advantages:
 - All the sequence information is used
 - Evaluate all possible trees
 - Sampling errors have least effect on the method
- Disadvantages:
 - Extremely slow
 - Impractical for analyzing large data set

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 evaluation

+ Bootstrapping & 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

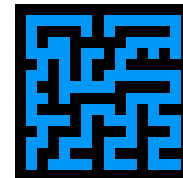
Tree Constructing Programs

- ✚ 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

The screenshot displays the ClustalX (1.83) software interface. The 'Trees' menu is open, showing options for alignment and tree construction. The main window shows a sequence alignment with 12 sequences, each 18 amino acids long. The alignment is color-coded by amino acid type: red for basic, green for acidic, blue for hydrophobic, yellow for polar, and purple for sulfur-containing. The sequences are:

```
* * : * * * : : : *****  
1 LI TRIMLIL SCICLASSLDGRFLAAAGIVV  
2 LI TRIMLTIGCIRFTGSLDGRFLAAAGIVV  
3 LI TRIMLILGCIRFTSSLDGRFLAAAGIVV  
4 LI TRIMLILGCIRFTSSLDGRFLAAAGIVV  
5 LI TRIMLILSCIHLTSSLDGRFLAAAGIVV  
6 LI TRIALALSCVHLASSLDGRFLAAAGIVV  
7 LI TRIMLILSCICPTGSLDGRFLAAAGIVV  
8 LTVRVALVLSVICPANSIDGRFLAAAGIVV  
9 LTVRVALVLSVICPANSIDGRFLAAAGIVV  
10 LTVRVALALSCVCPSSLDGRFLAAAGIVV  
11 LI TRIVALVLSVICPANSIDGRFLAAAGIVV  
12 LI TRIVALVLSVICPANSIDGRFLAAAGIVV
```

Tree Constructing Programs

File Edit Alignment Trees Colors Quality Help

Multiple Alignment Mode Font Size: 10

1	CH-A7/96	MGSTSSSTRIPAPPMLIITRIMLILSCICLASSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
2	Ch/98-3	MGSTSSSTRIPAPLMIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
3	Ch/99	MGSTSSSTRIPAPLMIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
4	Ch/2000	MGSTSSSTRIPAPLMIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
5	TW/2000	MGSRSSSTRIPAPLTIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
6	AUS	MGPRSSSTRIPAPLMIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
7	Ch/98-1	MGSTSSSTRIPVFPMLIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
8	Beaudette	MGPRPSTKNFVPMMLIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
9	V4_Queensland_/6	MGSRSSSTRIPVPLMIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
10	Ulster	MGSRSSSTRIPVPLMIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
11	Clone30	MGSRPSTKNFAPMMLIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL
12	LaSota	MGSRPSTKNFAPMMLIITRIMTLTGCIREETGSLDGRFLAAAGIVVTGDKAVNIYTSSQTGSIIVKLL

Output Format Options

CLOSE

Output Files

- CLUSTAL format
- NBRF/PIR format
- GCG/MSF format
- PHYLIP format
- GDE format
- NEXUS format
- FASTA format















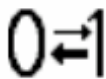



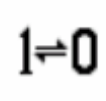


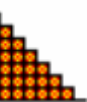
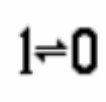
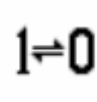
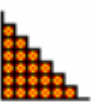
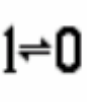
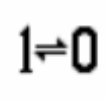









GDE output case : Lower

CLUSTALW sequence numbers : OFF

Output order : ALIGNED

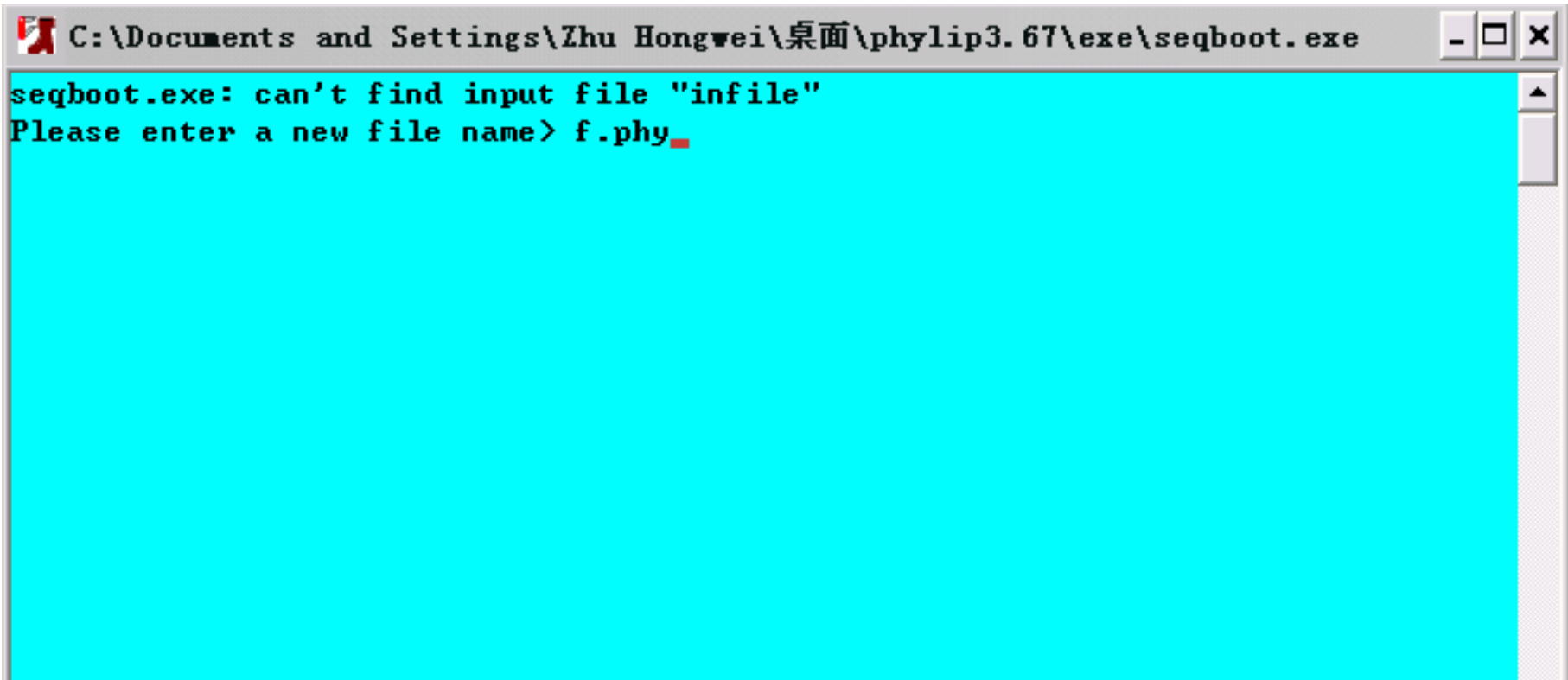
Parameter output : OFF

Tree Constructing Programs

	PHY 文件 9 KB		clique 应用程序		consense 应用程序		contml 应用程序
	contrast 应用程序		dnacomp 应用程序		dnadist 应用程序		dnainvar 应用程序
	dnaml 应用程序		dnamlk 应用程序		dnamove 应用程序		dnapars 应用程序
	dnapenny 应用程序		dollop 应用程序		dolmove 应用程序		dolpenny 应用程序
	drawgram 应用程序		drawtree 应用程序		factor 应用程序		fitch 应用程序
	gendist 应用程序		kitsch 应用程序		mix 应用程序		move 应用程序
	neighbor 应用程序		pars 应用程序		penny 应用程序		proml 应用程序
	promlk 应用程序		protdist 应用程序		protpars 应用程序		restdist 应用程序
	restml 应用程序		retree 应用程序		seqboot 应用程序		treedist 应用程序

Phylip program package

Tree Constructing Programs



A screenshot of a Windows command prompt window. The title bar shows the path `C:\Documents and Settings\Zhu Hongwei\桌面\phylip3.67\exe\seqboot.exe`. The window content is black with white text. The first line is an error message: `seqboot.exe: can't find input file "infile"`. The second line is a prompt: `Please enter a new file name> f.phy`, with a red cursor at the end of the text.

```
C:\Documents and Settings\Zhu Hongwei\桌面\phylip3.67\exe\seqboot.exe
seqboot.exe: can't find input file "infile"
Please enter a new file name> f.phy
```

Tree Constructing Programs

```
C:\Documents and Settings\Zhu Hongwei\桌面\phylip3.67\exe\seqboot.exe

Bootstrapping algorithm, version 3.67

Settings for this run:
  D      Sequence, Morph, Rest., Gene Freqs?  Molecular sequences
  J      Bootstrap, Jackknife, Permute, Rewrite?  Bootstrap
  %      Regular or altered sampling fraction?  regular
  B      Block size for block-bootstrapping?  1 <regular bootstrap>
  R      How many replicates?  100
  W      Read weights of characters?  No
  C      Read categories of sites?  No
  S      Write out data sets or just weights?  Data sets
  I      Input sequences interleaved?  Yes
  0      Terminal type <IBM PC, ANSI, none>?  IBM PC
  1      Print out the data at start of run  No
  2      Print indications of progress of run  Yes

  Y to accept these or type the letter for one to change
y_
```

Tree Constructing Programs

```
C:\Documents and Settings\Zhu Hongwei\桌面\phylip3.67\exe\seqboot.exe

Settings for this run:
D      Sequence, Morph, Rest., Gene Freqs?  Molecular sequences
J      Bootstrap, Jackknife, Permute, Rewrite? Bootstrap
%      Regular or altered sampling fraction?  regular
B      Block size for block-bootstrapping?  1 (regular bootstrap)
R      How many replicates?  100
W      Read weights of characters?  No
C      Read categories of sites?  No
S      Write out data sets or just weights?  Data sets
I      Input sequences interleaved?  Yes
0      Terminal type (IBM PC, ANSI, none)?  IBM PC
1      Print out the data at start of run  No
2      Print indications of progress of run  Yes

Y to accept these or type the letter for one to change
y
Random number seed (must be odd)?
23

completed replicate number  10
completed replicate number  20
completed replicate number  30
completed replicate number  40
completed replicate number  50
completed replicate number  60
completed replicate number  70
completed replicate number  80
completed replicate number  90
completed replicate number 100

Output written to file "outfile"

Done.
```

Tree Constructing Programs



Rename the 'outfile' to 'infile'

Tree Constructing Programs

```
C:\Documents and Settings\Zhu Hongwei\桌面\phylip3.67\exe\protdist.exe

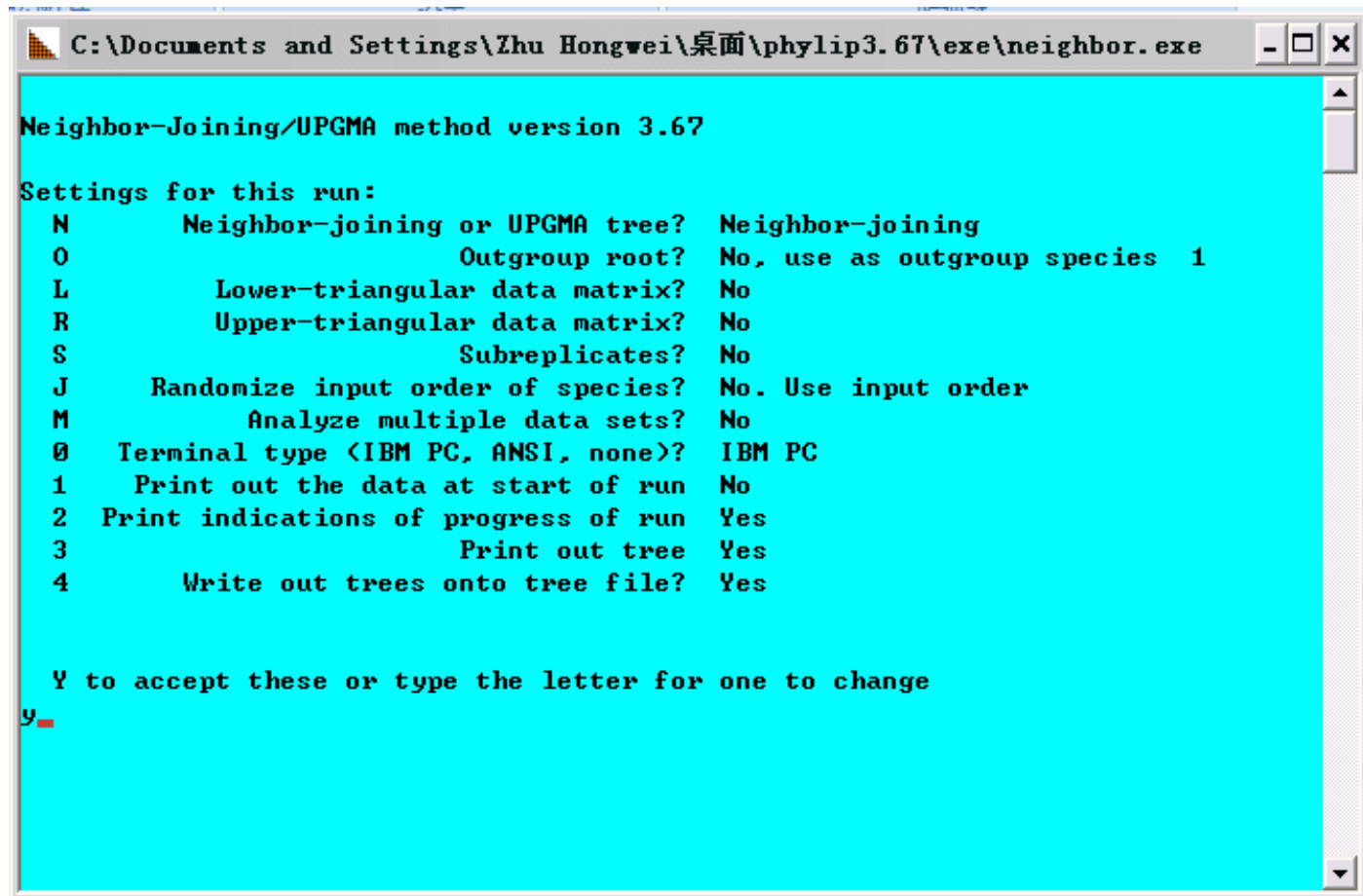
Protein distance algorithm, version 3.67

Settings for this run:
P Use JTT, PMB, PAM, Kimura, categories model? Jones-Taylor-Thornton matrix
G Gamma distribution of rates among positions? No
C      One category of substitution rates? Yes
W      Use weights for positions? No
M      Analyze multiple data sets? No
I      Input sequences interleaved? Yes
O      Terminal type (IBM PC, ANSI)? IBM PC
1      Print out the data at start of run? No
2      Print indications of progress of run? Yes

Are these settings correct? (type Y or the letter for one to change)
y_
```

Run 'protdist.exe' and an output of 'outfile'

Tree Constructing Programs



```
C:\Documents and Settings\Zhu Hongwei\桌面\phylip3.67\exe\neighbor.exe

Neighbor-Joining/UPGMA method version 3.67

Settings for this run:
N      Neighbor-joining or UPGMA tree?  Neighbor-joining
O      Outgroup root?  No, use as outgroup species  1
L      Lower-triangular data matrix?  No
R      Upper-triangular data matrix?  No
S      Subreplicates?  No
J      Randomize input order of species?  No. Use input order
M      Analyze multiple data sets?  No
0      Terminal type <IBM PC, ANSI, none>?  IBM PC
1      Print out the data at start of run  No
2      Print indications of progress of run  Yes
3      Print out tree  Yes
4      Write out trees onto tree file?  Yes

Y to accept these or type the letter for one to change
y_
```

Change 'outfile' to 'infile' and run 'neighbor.exe'

Tree Constructing Programs

```
C:\Documents and Settings\Zhu Hongwei\桌面\phylip3.67\exe\neighbor.exe

Y to accept these or type the letter for one to change
y

Cycle 9: species 7 < 0.00025> joins species 8 < 0.00318>
Cycle 8: node 7 < 0.01145> joins species 9 < 0.00929>
Cycle 7: species 10 < 0.00908> joins species 11 < 0.00297>
Cycle 6: node 7 < 0.03054> joins node 10 < 0.00822>
Cycle 5: node 7 < 0.00940> joins species 12 < 0.02431>
Cycle 4: species 6 < 0.04015> joins node 7 < 0.03101>
Cycle 3: species 4 < 0.00556> joins species 5 < 0.00818>
Cycle 2: node 4 < 0.00237> joins node 6 < 0.01347>
Cycle 1: species 1 < 0.00868> joins species 2 < 0.00335>
last cycle:
node 1 < 0.00013> joins species 3 < -0.00015> joins node 4 < 0.00971>

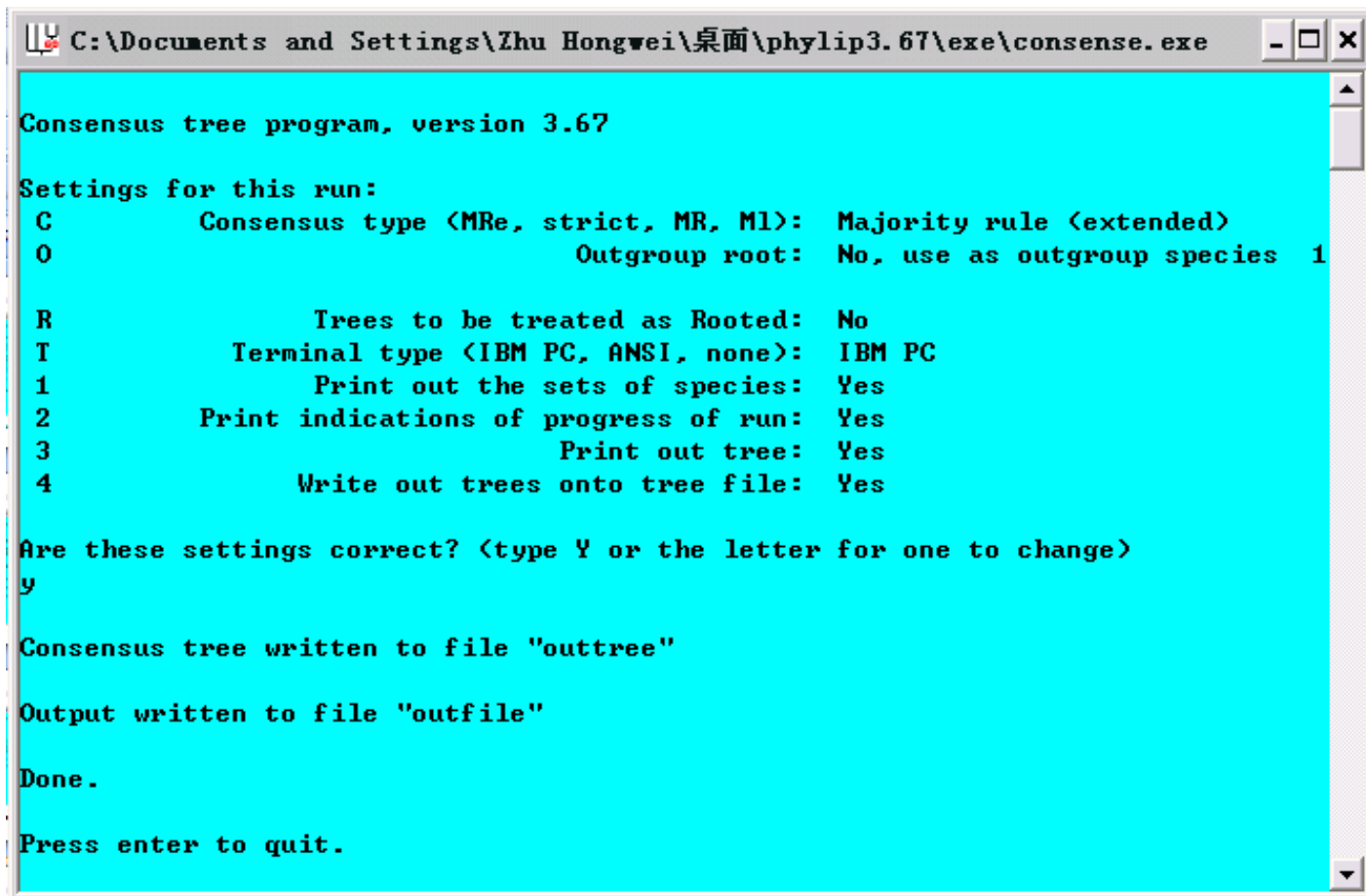
Output written on file "outfile"

Tree written on file "outtree"

Done.

Press enter to quit.
```

Tree Constructing Programs



```
C:\Documents and Settings\Zhu Hongwei\桌面\phylip3.67\exe\consense.exe

Consensus tree program, version 3.67

Settings for this run:
C      Consensus type (MRe, strict, MR, M1):  Majority rule (extended)
0      Outgroup root:                          No, use as outgroup species 1

R      Trees to be treated as Rooted:          No
T      Terminal type (IBM PC, ANSI, none):     IBM PC
1      Print out the sets of species:          Yes
2      Print indications of progress of run:   Yes
3      Print out tree:                         Yes
4      Write out trees onto tree file:         Yes

Are these settings correct? (type Y or the letter for one to change)
y

Consensus tree written to file "outtree"

Output written to file "outfile"

Done.

Press enter to quit.
```

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

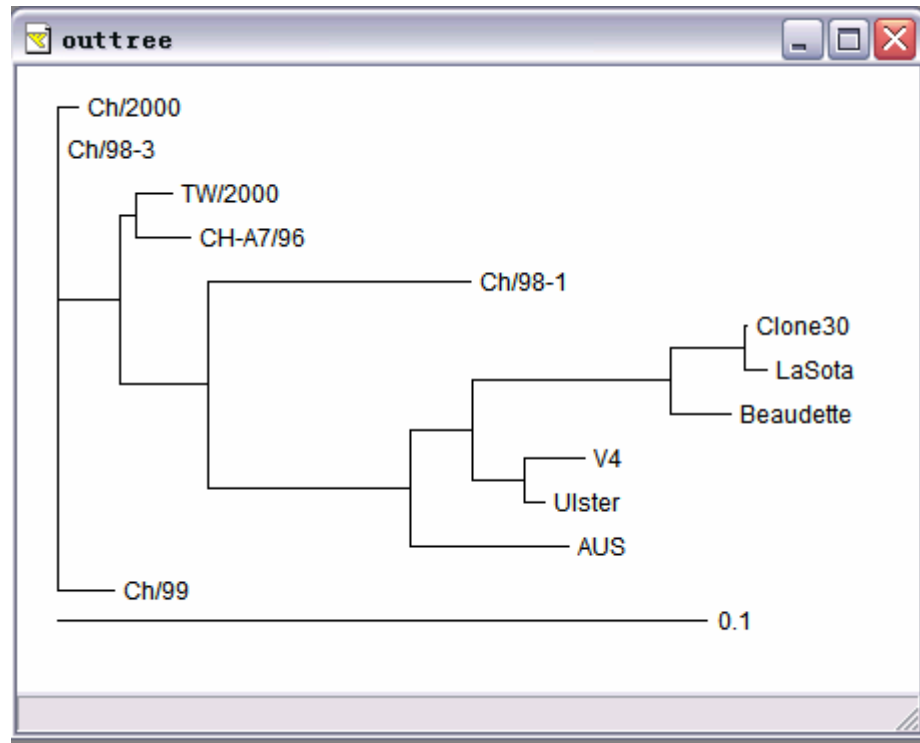
Tree Constructing Programs

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

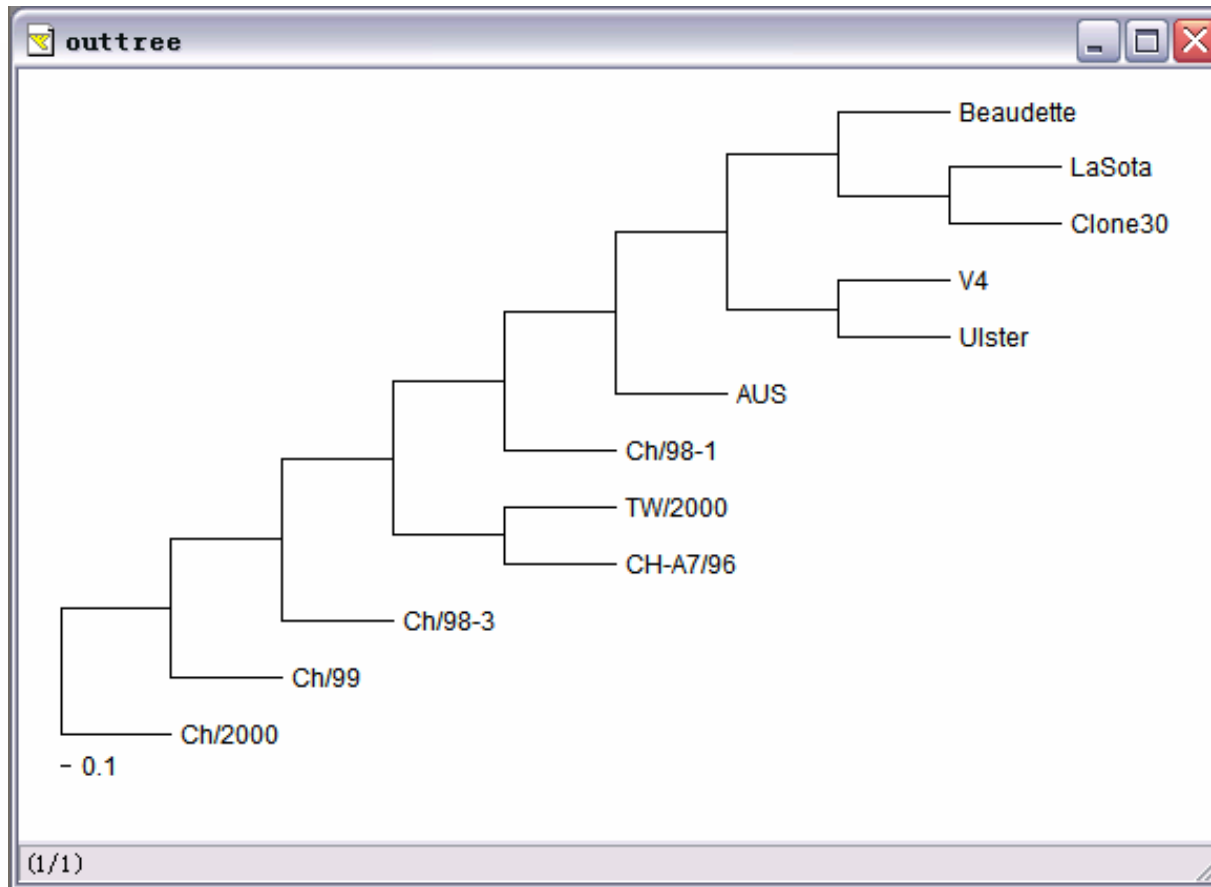
Neighbor-joining method

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
? ? ? ?
? ? ? ? +Beaudette
? ? ? ?
? ? ? ? +U4
? +-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!

!!